Aus der Klinik für Zahnerhaltung und Parodontologie

(Direktor: Prof. Dr. Christof Dörfer)
im Universitätsklinikum Schleswig-Holstein, Campus Kiel an der Christian – Albrechts-Universität zu Kiel

# Validierung von Vorhersagemodellen für Zahnverlust in einer Langzeitkohorte von Parodontitispatienten: Einfluss von Entwicklungs- und Validierungsstrategien.

Inauguraldissertation
zur
Erlangung der Doktorwürde der Zahnheilkunde
der Medizinischen Fakultät
der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Pia Antonia Brinkmann

aus Hamburg

Kiel (2020)

| | |
|---|---|
| 1. Berichterstatter/in: | Priv.-Doz. Dr. Christian Graetz |
| 2. Berichterstatter/in: | Priv.-Doz. Dr. Dr. Falk Birkenfeld |
| | |
| Tag der mündlichen Prüfung: | 07.12.2020 |
| Zum Druck genehmigt: | 07.12.2020 |

gez.: Prof. Dr. Johann Roider

Vorsitzender des Ausschusses für Promotion

# Inhaltsverzeichnis

# I. Abkürzungsverzeichnis

| | |
|---|---|
| **APT** | Aktive parodontale Therapie |
| **AUC** | Area under the Receiver Operating Characteristic Curve |
| **FN** | Falsch negative Ergebnisse |
| **FNR** | Falsch-negativ-Rate (False Negative Rate) |
| **FP** | Falsch positive Ergebnisse |
| **FPR** | Falsch-Positiv-Rate (False Positive Rate) |
| **GBM** | Gradient Boosting Machine |
| **KI** | Konfidenzintervall |
| **logR** | Logistische Regression |
| **OK** | Oberkiefer |
| **P** | Gesamtheit der tatsächlich positiven Ergebnisse |
| **P*** | Gesamtheit aller positiven Ergebnisse, korrekt positiv und falsch positiv |
| **RFO** | Random Forest |
| **ROC** | Operationscharakteristik eines Beobachters (Receiver Operating Characteristic) |
| **SD** | Standardabweichung (standard deviation) |
| **SHAP** | SHapley Additive exPlanations |
| **SHAP Value** | Shapley Value |
| **SMOTE** | Synthetic Minority Over-sampling Technique |
| **ST** | Sondierungstiefe |
| **TP** | Korrekt positive Ergebnisse (True Positive) |
| **TPR** | Richtig-positiv-Rate (True Positive Rate) |
| **TN** | Korrekt negative Ergebnisse (True Negative) |
| **TNR** | Richtig-negativ-Rate (True Negative Rate) |
| **TRIPOD** | Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis |
| **UK** | Unterkiefer |
| **UPT** | Unterstützende parodontale Therapie |

# 1. Einleitung

In der gegenwärtigen Individualmedizin nehmen Vorhersagemodelle eine zunehmend bedeutende Rolle ein. Sie befassen sich zum Beispiel mit der Vorhersage von Diabetes (Olivera et al., 2017) oder der Lebenserwartung nach einem ersten Myokardinfarkt (Wallert et al., 2017) und bilden die Grundlage einer präventiven Medizin der Zukunft. Modelle werden anhand von bekannten Daten (Trainingsdaten) entwickelt und trainiert, um appliziert auf unbekannte Daten (Testdatensatz) Vorhersagen treffen zu können. Von generalisierbaren, klinisch nützlichen Vorhersagemodellen verspricht man sich eine Steigerung der Effizienz der Behandlung und Objektivität für und bei der Prognosestellung (Kattan et al., 2013). Eine rigorose Entwicklung und Validierung von Vorhersagemodellen sollte stattfinden (Kundu et al., 2017). Hierbei müssen Entwickler beachten, dass eine vollständige Validierung eine interne Validierung zum Beispiel mittels Kreuzvalidierung und mehrere externe Validierungen („Narrow“- und „Broad“-Validierungen) beinhaltet, da nur so die Generalisierbarkeit der Modelle überprüft werden kann (Moons et al., 2015, Steyerberg and Harrell, 2016, Steyerberg et al., 2018). Eine „Narrow“-Validierung findet anhand von Daten aus einer ähnlichen klinischen Einrichtung oder Population statt, während eine „Broad“-Validierung anhand von Daten aus unterschiedlichen Einrichtungen mit variierender Prävalenz der Erkrankung stattfindet (Cowley et al., 2019). Für eine „Broad“-Validierung sind ebenfalls Daten, abhängig von der Länge des gesamten Beobachtungszeitraums, aus unterschiedlichen Zeitspannen („cross-temporal“) geeignet (Moons et al., 2015). Hinzu kommt die Wahl geeigneter Validierungsmetriken (Chawla, 2005, Perlich et al., 2003, Sun et al., 2011, Weiss, 2003). Neben der Diskriminations- soll auch die Kalibrierungsfähigkeit bestimmt werden (Cowley et al., 2019, Du et al., 2018, Vergouwe et al., 2002), welche aussagt, wie gut die vorhergesagten mit den tatsächlichen Befunden übereinstimmen. Eine Analyse über den klinischen Nutzen oder Einfluss des Modells auf Entscheidungen des Behandlers zum Beispiel anhand einer Decision-Curve-Analyse wird, sollte ein Modell bereit für eine Implementierung sein, nahe gelegt (Bouwmeester et al., 2012, Cowley et al., 2019, Vickers and Elkin, 2006).

Die Relevanz eines klinisch anwendbaren Vorhersagemodells für Zahnverlust wird anhand der Prävalenz von Parodontitis ersichtlich (Kassebaum et al., 2017). Sie ist die sechsthäufigste chronische Erkrankung weltweit (Kassebaum et al., 2014), welche zu

hohen gesundheitsökonomischen Kosten (Tonetti et al., 2017) führt. Der mit ihr oft in Verbindung stehende Zahnverlust, stellt den Behandler vor eine wichtige Entscheidung, da die Extraktion von Zähnen, wie auch ihre Retention verschiedene Risiken, Folgekosten sowie Einschränkungen für die Lebensqualität des Patienten (Ramseier et al., 2017) und die Prognosen der Nachbarzähne bergen kann. Aus der Natur der Daten zum Zahnverlust ergibt sich ein vieldiskutiertes (Ali, 2015, Chawla, 2005, García et al., 2007, Japkowicz, 2003, Sun et al., 2011), in der Zahnmedizin bis jetzt allerdings noch nicht adressiertes Problem. Es handelt von Daten mit einer ausgeprägten Unausgeglichenheit der Klassen, da Zahnverlust ein viel selteneres Ereignis im Vergleich zum Zahnerhalt darstellt. Dies wirft Probleme bezüglich des richtigen Klassifizierens der seltenen Klasse auf, da das Modell nur wenige Daten zum Erlernen des korrekten Vorhersagens von Zahnverlust gestellt bekommt und sich die Ausprägungen in einem Testdatensatz von diesen stark unterscheiden können. Tatsächlich ist es bei unausgeglichenen Daten meistens die seltene Klasse, die von besonderem Interesse ist, weshalb verschiedene Ansätze zum Umgang mit dieser Problemstellung vorgestellt wurden (Blagus and Lusa, 2013, Chawla, 2005, García et al., 2007, Japkowicz, 2003). So gibt es auf Datenebene verschiedene Under- und Over-sampling Techniken (s.u.), welche zum Angleichen der Klassen gedacht sind. Eine Möglichkeit besteht in der Anwendung von SMOTE (Synthetic Minority Over-sampling Technique, s.u.). Auf Algorithmus-Ebene soll das Erlernen zugunsten der Minoritätsklasse verstärkt werden. So kann beispielsweise eine Fehlprognose in dieser Klasse stärker gewichtet werden als in der Majoritätsklasse. Appliziert auf unterschiedliche Daten, führen diese Techniken zu kontroversen Ergebnissen. So gibt es keine Einstimmigkeit darüber, welche Technik zu empfehlen ist. Auf Daten zum Zahnverlust bei Parodontitispatienten wurde noch keine dieser Techniken angewandt.

## 1.1. Synthetic Minority Over-sampling Technique

SMOTE (Synthetic Minority Over-sampling Technique (Chawla and Hall, 2002)) ist eine Technik, die bei einer Ungleichverteilung von Klassen zur Anwendung kommt. Sie beinhaltet ein Over-sampling (zu Deutsch: Vermehrung von Stichproben) der Minoritätsklasse. SMOTE arbeitet auf der Ebene des Merkmalsraums und synthetisiert künstliche Beobachtungen in der k-Nearest-Neighborhood (zu Deutsch: „k-nächste-Nachbarschaft“) der seltenen positiven Beobachtungen (hier Zahnverlust). Die

Minoritätsklasse wird also mengenmäßig vergrößert. Je nach gewollter Over-sampling-Rate, werden k-Nearest-Neighbors zufällig gewählt und die Distanz zwischen diesen und der positiven Beobachtung berechnet. Multipliziert man diese Distanz mit zufälligen Zahlen zwischen 0 und 1, erhält man beliebig viele künstliche Beobachtungen, welche denen der Minoritätsklasse ähneln. Der Effekt ist, dass Entscheidungsräume (Klasse Plus) größer und weniger spezifisch werden. Die empfohlene Kombination mit Random Under-sampling (zu Deutsch: zufällige Reduktion von Stichproben), wo Beobachtungen bis zu einer vorher bestimmten Anzahl zufällig ausgewählt und entfernt werden, soll die anfänglichen Klassifikationsfehler zugunsten der Majoritätsklasse ausgleichen.

## 1.2. SHapley Additive exPlanations

Durch die zu erwartende Anwendung von Vorhersagemodellen im klinischen Alltag (American Academy of Periodontology, 2008), sollte eine einfache Interpretierbarkeit der Modellergebnisse gegeben sein, da sie den Behandler gerade bei komplexeren Modellen vor eine Herausforderung stellen. Der Grundgedanke ist es, komplexe Modelle anhand von vereinfachten, meist lokalen Erklärungsmodellen zugänglich zu machen. Wichtige Aspekte der Interpretierbarkeit sind, dass man erkennen kann, was das Modell zu einer Entscheidung bewegt hat, wie es zu einer bestimmten Prognose gekommen ist und ob dieser zu trauen ist.

Eine Visualisierungs- und Interpretationsmöglichkeit bietet SHAP, SHapley Additive exPlanations (Lundberg and Lee, 2017). SHAP kombiniert sechs bestehenden Methoden in einer additive feature attribution method (zu Deutsch: additive Merkmalszurechnungsmethode). Sie ordnen jedem Merkmal eine Einflussgröße auf das Ergebnis zu. Wendet man sie alleine an, weisen sie nicht die gleichen Eigenschaften wie in Kombination als SHAP auf. Nennenswert sind hier die lokale Genauigkeit (das lokale Erklärungsmodell muss mit dem eigentlichen Modell übereinstimmen), dass fehlenden Merkmalen kein Effekt zugeordnet wird und dass SHAP beständig ist. So sinkt die Einflussgröße einer Merkmalsausprägung nicht, wenn sie, ungeachtet der anderen Merkmale, durch Veränderungen in den Modellen einen gesteigerten Beitrag leistet (er wird nicht weniger wichtig). Außerdem muss sich der Behandler durch die Verwendung von SHAP nicht mehr entscheiden, welche Methode für seine Daten am besten geeignet ist. Dies ergibt sich nämlich aus dem von

Lundberg (2017) vorgestellten „unified framework“ (zu Deutsch: vereinheitlichtes Gerüst). Für alle möglichen Methoden liefert dieser eine einheitliche Lösung zur Erklärung eines jeden Ergebnisses eines Vorhersagemodells. Sogenannte Shapley Values bestimmen die Relevanz eines Merkmals für die Vorhersage anhand eines Vergleichs der Vorhersage unter und ohne Bezugnahme des Merkmals. Da die Reihenfolge der Bezugnahme von Merkmalen durch ein Modell Einfluss auf deren Effekt hat, wird der Vergleich in jeder möglichen Anordnung durchgeführt. Der SHAP Value kann lokal (Force Plot) für jedes Merkmal einer einzelnen Beobachtung oder global (Summary Plot) für die Merkmale aller Beobachtungen zusammen bestimmt werden. Im Sinne der „Game theory“ (zu Deutsch: Spieltheorie), nach deren Prinzipien sich SHAP orientiert, erhalten wir eine Belohnung in Form eines Ergebnisses und wissen, wie groß der Beitrag (SHAP Value) eines jeden Merkmals (Spielers) daran war. Um diesem Ansatz und dem zu interpretierenden Modell trauen zu können, muss abschließend jede Beobachtung validierbar sein. Ergibt sich kein zufriedenstellendes Ergebnis aus der Validierung, so muss das Modell noch verbessert werden. Hier findet sich der Nutzen von SHAP für die Modellentwicklung.

## 2. Hypothesen und Zielsetzungen der Arbeit

Im Rahmen dieser Studie wurden Modelle bezüglich ihrer Fähigkeit validiert, das Eintreten von Zahnverlust in einer Gruppe von Probanden mit bereits aktiv behandelter Parodontitis während der nachfolgenden unterstützenden parodontalen Therapie richtig zu prognostizieren. Es war nicht das Ziel dieser Arbeit, ein klinisch anwendbares Modell zu entwickeln, sondern vielmehr sollen die Ergebnisse und Implikationen dieser Validierungsstudie als weiterer Baustein für die Evaluation von Entwicklungs- und Validierungsstrategien für Vorhersagemodelle genutzt werden. Langfristig kann so eine rigorose Entwicklung generalisierbarer und verlässlicher Modelle ermöglicht werden und der am besten zu der Problemstellung passende Algorithmus identifiziert werden. Im Vorfeld wurden in der standortübergreifenden Arbeitsgruppe unter Leitung von Prof. Dr. Schwendicke (Abteilung für Zahnerhaltung und Präventivzahnmedizin, Charité – Universitätsmedizin Berlin) und PD Dr. Graetz (Klinik für Zahnerhaltung und Parodontologie, Universitätsklinikum Schleswig-Holstein, Campus Kiel) bereits Analysen mit verschiedenen Datensätzen durchgeführt, welche für komplexere Modelle (Random Forest (RFO), Extreme Gradient Boosting Machine) keine Vorteile

gegenüber der weniger komplexen, wie der logistischen Regression (logR) zeigten. Aufbauend auf diese Erkenntnisse, untersucht die vorliegende Arbeit spezifisch nur die Langzeitdaten einer Kieler-Kohorte von Patienten mit Parodontitis, behandelt in der Klinik für Zahnerhaltung und Parodontologie des Universitätsklinikum Schleswig-Holsteins.

Die primäre Hypothese lautet, dass die AUC (Area under the Receiver Operating Characteristic Curve) der Modelle mit RFO und GBM (Gradient Boosting Machine) in der In-sample-Validierung signifikant größer sind als die der logR. Als Nullhypothese wird angenommen, dass die AUC der Modelle mit RFO und GBM nicht signifikant höher sind als die der logR.

Im Rahmen der Studie wurde erstmals SMOTE auf die Daten von Zahnverlustmodellen angewandt und es gilt zu prüfen, ob diese Technik die Vorhersagekraft, gemessen an der AUC, der Modelle verbessert. Dabei wird hypothetisch angenommen, dass unter der Verwendung von SMOTE die Ergebnisse der AUC von komplexeren Modellen (RFO, GBM) in der In-sample-Validierung signifikant höher sind als ohne die Verwendung von SMOTE. Außerdem ermöglicht das Durchführen einer ersten externen Validierung, mittels „Narrow“-Validierung, einen Vergleich der Ergebnisse der In-sample und der Out-of-sample Validierung. Dadurch soll eine Einschätzung der Generalisierbarkeit ermöglicht werden. Diese spezifischen Erkenntnisse der Untersuchung könnten genutzt werden, um entgegen des zu verzeichnenden Analysetrends, bestehende Erkenntnisse zur Entwicklung und Validierung außer Acht zu lassen, diese stattdessen in die Modelle mit einzubeziehen. Dadurch soll es für zukünftige Forschungsprojekte möglich werden, die beeinflussenden Entwicklungsschritte besser zu verstehen und mögliche neue Implikationen für eine erweiterte Validierung zu berücksichtigen.

## 3. Forschungsstand, Literaturrecherche

Vorhersagemodelle in der Zahnmedizin werden entwickelt, um Prognosen über das Auftreten von Parodontitis und ihre Komplikationen, wie Zahnverlust, stellen zu können.

Es wurden verschiedene Vorhersagemodelle vorgestellt. Sie erzielen oft nur unzureichende Ergebnisse, was vor allem auf Entwicklungs-, Berichterstattungs- und Validierungsfehler zurückzuführen ist (Bouwmeester et al., 2012, Du et al., 2018,

Moons et al., 2015, Schwendicke et al., 2018, Steyerberg and Vergouwe, 2014, Steyerberg et al., 2018). So berichteten Du et al.(Du et al., 2018), dass bestehende Leitlinien oftmals nicht mit einbezogen werden und die Validität, die praktische Anwendbarkeit (Schwendicke et al., 2018) sowie die Existenz eines klinisch anwendbaren Vorhersagemodells in Frage zu stellen sind. In ihrer Studie untersuchten sie fünf Vorhersagemodellstudien. In den vier Entwicklungs- und der einen Validierungsstudie fand keine interne (Kreuzvalidierung) und nur einmalig eine externe Validierung statt. Dies führte zu Overfitting (zu Deutsch: Überanpassung) und einer Überschätzung der Modellleistung. Die Kalibrierungsfähigkeit wurde in keiner Studie bestimmt. Die Leistung der Modelle ist hier nicht vergleichbar, da die Studien alle unterschiedlich aufgebaut sind, unterschiedliche Ergebnisse gesucht werden, die Beobachtungsspanne variiert, verschiedene Validierungsmetriken verwendet werden oder die Validierungen mangelhaft, die Validität also als fragwürdig einzustufen ist. In der Studie von Schwendicke et al. (2018) rangieren die AUC-Werte der untersuchten Klassifikationsmodelle (Avila et al., 2009, Fardal et al., 2004, Nunn et al., 2012), getestet an einem Kieler Datensatz mit 301 Probanden, zwischen 0,52 und 0,67. Dabei entspricht ein AUC-Wert von 0,5 dem Zufall. Am besten schneidet hier das Klassifikationsmodell von Avila et al. (2009) mit einer AUC von 0,67 (KI: 0,65/0,69) ab. All diese getesteten Modelle befassen sich mit der Risikobewertung für Zahnausfall und teilen die Beobachtungen in Klassen ein. So werden in dem Modell von Avila et al. (2009) die Zähne in fünf Grade von *Zahnerhalt empfohlen* bis *Extraktion empfohlen* eingeteilt. Die Diskriminationsfähigkeit ist moderat. Das bereits erwähnte Problem der Klassenunausgeglichenheit spiegelt sich hier in einer hohen Spezifität (84%) gegenüber einer niedrigen Sensitivität (46%) wider. Es wird nicht einmal die Hälfte der Zahnausfälle (positive Klasse) korrekt prognostiziert, während 84% der Zähne richtig in die Klasse „Retention“ (kein Zahnverlust, negative Klasse) einsortiert werden. Ähnliche Ergebnisse finden sich ebenfalls bei anderen Modellen dieser Art (Martinez-Canut et al., 2017). Bei sehr guten Spezifitätswerten, erreicht die Sensitivität bei eigener Validierung durch die jeweiligen Autorengruppen nur Werte um die 40%. Die Diskriminationsfähigkeit eines Modells wird bei stark unausgeglichenen Klassen zugunsten der negativen, häufig auftretenden Klasse verschoben (Chawla, 2005). Dies lässt sich an der sehr geringen Anzahl an positiven Beobachtungen zum Trainieren der Modelle erklären. Die Modelle haben kaum Material, um das richtige Prognostizieren der positiven Klasse (entspräche einer hohen Sensitivität) zu erlernen

und die Modelle erwarten eher, dass Zähne erhalten bleiben. Die Autoren helfen sich in dem Fall von Martinez-Canut et al. (2017) mit der Aussage, dass das Modell zur Vorhersage für einen nicht eintretenden Zahnausfall verwendet werden kann und dass ein falsch positives Ergebnis, ein Zahn fällt trotz Vorhersage nicht heraus, weniger schlimm sei als ein falsch negatives Ergebnis, ein nicht vorhergesagter Zahnverlust tritt ein. Diese Annahme passt zu dem seltenen Auftreten von Zahnverlust und lässt die Ergebnisse besser aussehen, ändert aber nichts an der Tatsache, dass das Modell hinsichtlich seines Entwicklungsziels Zahnverlust vorherzusagen, nur moderat abschneidet. Die Autoren sehen ihr Modell selber nur als Ansatz. Allgemein zeigen diese Ergebnisse trotz der mangelnden Vergleichbarkeit, dass es den Modellen an Genauigkeit, Reliabilität und Validität fehlt.

# 4. Material und Methodik

Diese Studie richtet sich nach den TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) Berichterstattungsrichtlinien (Moons et al., 2015).

Drei unterschiedlich komplexe Modelltypen (logistische Regression, Random Forest und Gradient Boosting Machine) werden anhand einer Langzeitkohorte von Parodontitispatienten, welche seit 1982 in dem Funktionsbereich Parodontologie der Klinik für Zahnerhaltung und Parodontologie im Universitätsklinikum Schleswig-Holstein behandelt werden, validiert. Wie bereits erwähnt, war es nicht das Ziel dieser Arbeit, klinisch anwendbare Modelle zu entwickeln. Sondern es sollte überprüft werden, inwieweit die Vorhersagen durch die jeweiligen Modelle mit den tatsächlichen klinischen Befunden während der UPT übereinstimmen und die Diskriminationsstärken der einzelnen Modelle sollten evaluiert beziehungsweise verglichen werden.

Aufgrund der Klassenunausgeglichenheit wurde auf jedes Modell in einem zusätzlichen Schritt SMOTE angewendet und bestimmte Validierungsmetriken erhoben. Die deskriptive statistische Evaluation wurde mittels Phyton’s scientific stack (Python Software Foundation, https://www.python.org/) und Jupyter Notebooks (Project Jupyter, https://jupyter.orgin) in der Charité, Universitätsmedizin Berlin von Herrn Dr. Krois durchgeführt. Tabellen und Abbildungen wurden teils mit Excel für Mac 2019 (Microsoft, Redmond, WA, USA) erstellt. Ein Ansatz für eine einfachere

Interpretierbarkeit von Vorhersagemodellergebnissen im klinischen Alltag wurde durch die Verwendung von SHAP (Shapley Additive exPlanations) vorgestellt und anhand der Ergebnisse diskutiert.

## 4.1. Datenquelle, Teilnehmer, Einschlusskriterien und Therapiekonzept

Die im vorherigen Absatz genannte Patientenkohorte war bereits mehrfach Gegenstand von longitudinalen Studien (Graetz et al., 2011, Graetz et al., 2017a, Graetz et al., 2017b). Die Einschlusskriterien in diese Kohorte waren, dass Patienten eine chronische oder aggressive Parodontitis haben müssen, wobei die chronische Parodontitis nach der Klassifikation von 1999 (Armitage, 1999) definiert war, während die Kriterien für eine aggressive Parodontitis modifiziert wurden. Es wurden ausschließlich Patienten mit einem Knochenabbau von ≥ 50% an ≥ zwei Zähnen (Kim et al., 2006)) nach Abschluss der aktiven parodontalen Therapie (APT (T0-T1)) und nach ≥ neun Jahre lang durchgeführter unterstützender parodontaler Therapie (UPT (T1-T2)) mit ≥ einem Termin pro Jahr mit jährlicher Befunderhebung der Sondierungstiefen und einer radiologischen Dokumentation zum Zeitpunkt T0 und dem letzten Termin im Rahmen der UPT (T2) aufgenommen. 390 Patienten entsprachen diesen Kriterien. Patienten mit unvollständiger Sondierungstiefen- oder röntgenologischer Dokumentation wurden nicht mit aufgenommen (n=2174). Die Patienten erhielten im Rahmen der Behandlung im Funktionsbereich Parodontologie im Universitätsklinikum Schleswig-Holstein eine aktive und unterstützende parodontale Therapie. Dieses Behandlungskonzept wurde bereits ausführlich andernorts beschrieben (Graetz et al., 2011, Graetz et al., 2017b). Erwähnt sei aber, dass in der APT hauptsächlich nicht-operativ vorgegangen wurde und es neben einem mechanischen geschlossenen Vorgehen (Subgingivales Scaling und Root planing) nur mit bestimmter Indikation zu einer unterstützenden offenen Zugangslappenoperation kam. Im Fall einzelner Patienten wurden endodontische Maßnahmen ergriffen, bewegliche Zähne geschient, mehrwurzelige Zähne tunneliert, gesteuerte Geweberegeneration wurde angewandt oder es wurden Wurzelresektionen durchgeführt. Operationen zur Taschentiefenreduktion, Knochenresektion oder -augmentation waren nicht Teil der APT in Kiel. Eine unterstützende systemische Antibiose in Form des van Winkelhoff Cocktails (Winkelhoff et al., 1996) fand nur bei Patienten statt, bei denen trotz adäquater Therapie und guter Mundhygiene die

Parodontitis nach erneuter Desintegration des Biofilms fortschritt. Nach Abschluss der APT folgte die UPT in individuellen Abständen von drei bis zwölf Monaten. In deren Rahmen wurde regelmäßig bezüglich der individuellen Mundhygiene re-instruiert und -motiviert sowie eine Zahnreinigung durchgeführt. Bei Bedarf wurden Resttaschen entweder nicht-chirurgische subgingival nachinstrumentiert oder bei entsprechender Indikation unter Sicht im Rahmen von Zugangslappenoperationen behandelt. Die Ethikkommission der Christian-Albrechts-Universität zu Kiel gab im Rahmen der longitudinalen Nachuntersuchungen ein positives Votum zur Durchführung (D489/13), welches 2017 durch ein Amendement für weitere Fragestellungen ergänzt wurde (Kiel: 566/17).

## 4.2. Zielvariable

Das gesuchte Ergebnis war Zahnverlust während der UPT (T1-T2). Hierbei wurden nicht nur Zahnverluste mit einbezogen, die sich auf Parodontitis zurückführen ließen, sondern ebenfalls solche, die sich aus unbekannten oder multiplen Gründen ereigneten. Es ist davon auszugehen, dass Zähne nicht unbedacht extrahiert wurden. Trotzdem muss gerade bei den Extraktionsfällen, welche außerhalb des Funktionsbereiches stattfanden, bedacht werden, dass in Einzelfällen auch bestimmte Ziele mit der Extraktion verfolgt wurden (z.B. finanzielle Interessen, Praxisrelevante Konzepte etc.), welche retrospektiv nicht abgeschätzt werden können.
Zahnverluste während der APT wurden gar nicht mit einbezogen, da die Extraktionsgründe neben einem aussichtslosen Zahnerhaltungsversuch trotz adäquater Therapie (nicht erhaltungsfähiger Zahn) auch häufig aufgrund einer Therapieplanung mit umfassender oraler Rehabilitation des Gesamtgebisses stattfanden (kassenzahnärztliche Richtlinien, Pfeilerwertigkeit etc.).

## 4.3. Prädiktoren

Von den Patienten der Langzeitkohorte standen drei patientenbezogene und fünf zahnbezogene Prädiktoren zur Verfügung. Diese wurden konsequent einmal jährlich erhoben. Manche von ihnen stammen ausschließlich aus T0 und nicht aus T1, zumeist begründet in Ermangelung einer Indikation für weitere bildgebende Verfahren mit Röntgenstrahlen (Strahlenschutzverordnung). Somit kann angenommen werden, dass

keine signifikanten Veränderungen in dieser Zeitspanne auftraten, welche zusätzliche Röntgenbilder erforderlich gemacht hätten. Die patientenbezogenen Prädiktoren sind:

- Patientenalter und Geschlecht zum Zeitpunkt T0,
- Anzahl an Zähnen nach Abschluss der APT,
- Raucherstatus in Kategorien nach Lang und Tonetti (2003) nie/ zuvor (>5Jahre her)/ aktuell) zum Zeitpunkt T0.

Folgende zahnbezogene Prädiktoren standen aufgrund der jährlich erhobenen Befunde zur Verfügung (Daten der dritten Molaren wurden exkludiert):

- Typ des Zahns (Molar oder Nicht-Molar),
- maximalste Tiefe der Sondierungstiefen (ST) eines Zahnes,
- Zahnbeweglichkeit nach den Kategorien von (Lindhe and Nyman, 1977) in Grad 1-3 (Grad 0 definiert als keine Zahnbeweglichkeit),
- Relativer röntgenologischer Knochenabbau (%) zum Zeitpunkt T0, kategorisiert nach ≤25%, >25-50%, >50-70% und >70% (Graetz et al., 2011),
- Furkationsbeteiligung von Molaren nach (Hamp et al., 1975) an T1. Es wird ausschließlich die höchste gemessene Furkationsbeteiligung angegeben (McGuire, 1996).

## 4.4. Statistische Experimente

Verschiedene Experimente wurden durchgeführt, um den Einfluss der Komplexität der Modelle auf ihre Vorhersagekraft und ihre Generalisierbarkeit (1) ohne und (2) mit SMOTE bei gleichbleibender Validierungsstrategie zu bestimmen. Alle Modelle wurden intern (In-sample) mittels fünffacher Kreuzvalidierung und im Rahmen der Studie extern (Out-of-sample) validiert.

(1) Komplexität ohne SMOTE: Verschiedene Metriken wurden zur Bestimmung der Vorhersagekraft der verschieden komplexen Modelle erhoben (s.u.). Einseitige, binomiale Hypothesentests wurden durchgeführt, um zu prüfen, ob die AUC-Werte der komplexen Modelle in der In-sample-Validierung, denen der logR signifikant unterlegen waren.

(2) Applikation von SMOTE: Over-sampling der nur geringfügig vertretenden Klasse „Zahnverlust“ durch Generierung künstlicher Beobachtungen. Für einen Vergleich mit und ohne Anwendung der Technik wurden die

Validierungsmetriken erhoben und ein einseitiger, binomialer Hypothesentest wurde mit den Ergebnissen der In-sample-Validierung durchgeführt, um zu prüfen, ob die AUC-Werte mit SMOTE niedriger waren als ohne.

(3) Die Metriken der Out-of-sample-Validierung wurden in einem weiteren Schritt mit denen der In-sample-Validierung verglichen.

### 4.4.1. Untersuchte Modelle

Es wurden drei verschieden komplexe Modelle (definiert nach der Anzahl von Hyperparametern) verwendet. Nach ihrer Komplexität sortiert, handelte es sich um eine logistische Regression (logR), einen Random Forest (RFO) und eine Gradient Boosting Machine (GBM).

Eine logistische Regression ist ein oftmals angewandtes Modell zur Modellierung dichotomer Zielvariablen. Das Modell gehört der Klasse der „Generalized Linear Models“ (Wedderburn, 1972) an, also einem verallgemeinerten linearen Modell und wendet eine Logit-Kopplungsfunktion an, um die Beziehung zwischen den Prädiktoren und der binären Antwortfunktion modellieren zu können, welche sich in Wahrscheinlichkeiten konvertieren lässt.

Bei einem Random Forest handelt es sich um ein Klassifikationsverfahren, welches aus mehreren nicht korrelierten Entscheidungsbäumen besteht. Er gehört zu der Klasse der Ensemblemethoden, bei denen mehrere Modelle (Algorithmen) kombiniert werden, um gemeinsam ein Problem zu lösen (Breimann, 2001). Hierbei kann es sich um Regressions- wie auch um Klassifikationsfälle handeln. Zur Konstruktion der Sammlung von Modellen (Entscheidungsbäumen) kombiniert ein RFO Bagging (von engl. Bootstrap Aggregating) und eine randomisierte Variablenauswahl. Bagging bedeutet, dass das Trainingsdatenset in kleine Sets aufgeteilt wird und für jedes dieser Sets ein eigenes Modell angepasst wird. Kombiniert mit der randomisierten Variablenauswahl, weisen die einzelnen Modelle unterschiedliche Variablen auf. Jedes dieser Modelle liefert eine eigene Vorhersage für einen unbekannten Datenpunkt, welche im Klassifikationsfall als Stimme betrachtet wird. Die finale Klassifizierung entspricht dann der Klasse mit den meisten Stimmen.

Eine GBM gehört ebenfalls zu den Ensemblemethoden, die auf Entscheidungsbäumen basieren und kann für Klassifikations-, Regressions- und Rankingfälle verwendet werden. Im Klassifikationsfall werden durch sogenanntes

Boosting die Vorhersagen einer großen Anzahl an schwachen Klassifikatoren kombiniert. Obwohl die schwachen Klassifikatoren kaum genauer sind, als der Zufall, so können diese Klassifikatoren, appliziert auf immer wieder modifizierte Versionen der Daten, sich zu starken Ensemblen, zu einem „Committee“, entwickeln (Hastie, 2009). Die Idee zur Verbesserung des Modells ist es, dieses schrittweise im additiven Sinn zu verändern, so dass folgende Klassifikatoren so trainiert werden, dass sie Beobachtungen richtig klassifizieren können, die ihre Vorgänger noch falsch zuordneten.

### 4.4.2. Modellierungs- und Validierungsmetriken

Es wurden die " Area under the Receiver Operating Characteristic Curve" (AUC), die Accuracy und ihr 95% Konfidenzintervall (KI), die Sensitivität, die Spezifität und der F1-Score bestimmt. Ein Klassengrenzwert von 0,5 wurde angenommen. Die Baseline Accuracy wurde außerdem mit der Accuracy der applizierten Modelle verglichen, da letztere bei nützlichen Modellen höher sein sollte, als bei der ausschließlichen Vorhersage der häufigeren Klasse.

Für aussagekräftige Ergebnisse bezüglich der Vorhersagen teilt man diese in eine Konfusionsmatrix auf (Abb.1).

| | | |
|---|---|---|
| Tatsächlich negativ | Richtig negativ (TN) | Falsch positiv (FP) |
| Tatsächlich positiv | Falsch negativ (FN) | Richtig positiv (TP) |
| | Vorhersage negativ | Vorhersage positiv |

a

| | | |
|---|---|---|
| Tatsächlich negativ | Richtig-negativ-Rate (TN) | Falsch-positiv-Rate (FP) |
| Tatsächlich positiv | Falsch-negativ-Rate (FN) | Richtig-positiv-Rate (TP) |
| | Vorhersage negativ | Vorhersage positiv |

b

Abbildung 1a und b: Aufbau einer Konfusionsmatrix. 1a: Ohne Normierung. Sie beinhaltet als Elemente die Anzahl an richtig positiven (True Positives, TP), richtig negativen (True Negatives, TN), falsch positiven (False Positives, FP) und falsch negativen (False Negatives, FN) Prognosen.1b: Normiert stellen sie die Richtig-positiv-Rate (TPR, True Positive Rate), die Falsch-positiv-Rate (FPR, False Positive Rate), die Richtig-negativ-Rate (TNR, True Negative Rate) und auch die Falsch-negativ-Rate (FNR, False Negative Rate) dar.

Aus den verschiedenen Funktionen der Konfusionsmatrix ergeben sich folgende Metriken:

1. *Sensitivität*

   Die Sensitivität, auch Recall oder TPR genannt, berechnet sich aus den korrekt als positiv klassifizierten Ergebnissen (TP) in Relation zu der Gesamtheit der tatsächlich positiven Ergebnisse (P). Also tatsächlicher Zahnverlust, der auch als solcher vorhergesagt wurde.

$$Sensitivität = \frac{TP}{TP + FN} = \frac{TP}{P}$$

2. *Spezifität*

Die Spezifität oder TNR hingegen berechnet sich aus dem Anteil der korrekt als negativ prognostizierten Ereignisse (TN) in Relation zu den tatsächlich negativen (N).

$$Spezifität = \frac{TN}{TN+FP} = \frac{TN}{N}$$

3. *Accuracy* (zu Deutsch: Korrektklassifikationsrate)

   Die Accuracy beschreibt den Anteil der korrekten Prognosen von der Grundgesamtheit. Für stark unausgeglichene Klassen ist diese Metrik nicht geeignet (Chawla, 2005, Perlich et al., 2003), da Modelle beispielsweise anhand von Daten mit einer Häufigkeitsverteilung der Klassen von 1/10 bereits eine Accuracy von 90% durch das einfache Vorhersagen der Majoritätsklasse erreichen würden.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

4. *F1-Score*

   Der F1-Score fasst die Metriken Precision (auch Positive Predictive Value, zu Deutsch: positiver Vorhersagewert genannt) und Recall (Sensitivität) zusammen. Es handelt sich hierbei um das harmonische Mittel der richtig vorhergesagten positiven Prognosen (TP) in Relation zu der Grundgesamtheit

aller positiven Ergebnisse (P*) und der korrekt als positiv klassifizierten Ergebnisse (TP) in Relation zu den tatsächlich positiven Ergebnissen (P).

$$F1 = \left( \frac{\frac{1}{Precision} + \frac{1}{Recall}}{2} \right)^{-1}$$

mit

$$Precision = \frac{TP}{TP+FP} = \frac{TP}{P^*}$$

und

$$Recall = \frac{TP}{TP + FN} = \frac{TP}{P}$$

### 5. *ROC/AUC*

Eine weitere, häufig verwendete Metrik sind ROC-Kurven **(**ROC: Englisch für Receiver Operating Characteristic) auch Grenzwertoptimierungskurven genannt, in denen die Sensitivität für verschiedene Tests gegen die FPR (1-Spezifität) aufgetragen wird. Die Kurve bildet relative Trade-offs zwischen Nutzen (TP) und Kosten (FP) für verschiedene Grenzwerte der Vorhersagemodelle ab. Die Fläche unter dieser Kurve bezeichnet man als AUC (Area under the Receiver Operating Characteristic Curve). Ein optimales Modell würde einen Wert von 1 erhalten, 0,5 entspricht dem Zufall. Die ROC ist unabhängig von den Proportionen der Klassen (Menon et al., 2013) und damit als Performancemetrik für Modelle, die mit unausgeglichenen Daten arbeiten, geeignet. Die ROC-Kurven der Modelle sind in Abbildung 3 dargestellt.

## 4.5. Entwicklungsdaten vs. Validierungsdaten

Die Datensätze stammen aus der gleichen Langzeitkohorte mit gleichen Aufnahmekriterien und weisen abgesehen von ihrer Stichprobengröße (5697 vs. 1813 Zähne) keine Unterschiede bezüglich ihres gesuchten Ausgangs und ihrer gewählten Prädiktoren auf. Das Ereignis Zahnverlust macht jeweils in etwa 10% der Daten aus, wobei die Wahl der entsprechenden Beobachtungen zufällig ablief.

# 5. Ergebnisse

## 5.1. Stichproben und Zahnverlust

Die Eigenschaften der Langzeitkohorte sind in Tabelle 1 dargestellt. Sie umfasst 390 Patienten. Das durchschnittliche Alter (SD) zum Zeitpunkt T0 betrug 45,9 (10,2) Jahren. Der Mittelwert des Zahnverlusts betrug 0,11 (0,15) Zähne pro Patienten und Jahr.

| **Parameter** | |
|---|---|
| Anzahl Patienten (männlich/ weiblich) | 164/226 |
| Patientenalter zum Zeitpunkt T0 (Mittelwert, SD) in Jahren | 45.9 (10.2) |
| UPT (T1-T2) in Jahren (365 Tage) (Mittelwert, SD) | 18.2 (5.6) |
| Anzahl an Rauchern/ Ex-Rauchern/ Nichtraucher (T0) | 50/ 88/ 252 |
| Anzahl von Zahnverlust/ (Patient*Jahr) (UPT) (Mittelwert, SD) | 0.11 (0.15) |

Tabelle 1: Charakteristika der Stichproben zu unterschiedlichen Zeitpunkten. Phase der aktiven Parodontitistherapie (APT) von T0-T1 (erster bis letzter APT Besuch), Phase der unterstützenden Parodontitistherapie (UPT) von T1-T2 (letzter APT Besuch bis zum letzten UPT Besuch). N=390 Patienten, 7518 Zähne.

Zahnverlust trat in erster Linie bei älteren Patienten und noch öfter bei solchen, die rauchten auf. Aus zahnbezogener Sicht gingen Molaren, Zähne mit hohen Sondierungstiefen (ST), Zähne mit Furkationsbeteiligung und Zähne mit starkem Knochenabbau und hoher Beweglichkeit am häufigsten verloren (Tabelle 2). Da das Aufteilen der Daten zufällig unter Beachtung einer annähernd gleichen Wahrscheinlichkeitsverteilung der Ereignisse erfolgte, wurden keine Unterschiede zwischen den Datensätzen erwartet.

| ***Patientenbezogen*** | *Patienten mit Zahnverlust/ Total (%)* | ***Zahnbezogen*** | *Zahnverlust/ Total (%)* |
|---|---|---|---|
| ***Alter T1***, *(Mittelwert, SD)* | *Verlust: 46,7 (9,4) Retention:44,7(11,3)* | **Zahntyp** | |
| | | Molar | 392/2060 (19,2%) |
| | | Nicht-Molar | 334/5450 (6,1%) |
| ***Raucherstatus*** | | **Sondierungstiefen (ST)** | |
| *Nie* | *155/252 (61,5%)* | <5mm | 466/6439 (7,2%) |
| *Zuvor* | *54/88 (61,4%)* | 5-7mm | 228/1001 (22,8%) |
| *Raucher* | *36/50 (72,0%)* | >7mm | 32/70 (45,7%) |
| ***Geschlecht*** | | **Furkationsbeteiligung** | |
| *Männlich* | *103/164 (62,8%)* | Mit | 142/378 (37,6%) |
| *Weiblich* | *142/226 (62,8%)* | Ohne | 584/7132 (8,2%) |
| | | **Knochenabbau** | |
| | | ≤ 25% | 80/2191 (3,7%) |
| | | 25-50% | 237/3316 (7,1%) |
| | | 50-70% | 290/1665 (17,4%) |
| | | >70% | 119/338 (35,2%) |
| | | **Beweglichkeit** | |
| | | 0 | 591/7124 (9,0%) |
| | | 1 | 95/288 (33,0%) |
| | | 2 | 21/63 (35,4%) |
| | | 3 | 19/37 (53,8%) |

Tabelle 2: Verteilung des Zahnverlustes in Bezug auf verschiedene patienten- und zahnbezogene Variablen aus den gesamten zur Verfügung stehenden Daten (n=390 Patienten, 7518 Zähne).

## 5.2. Einfluss der Komplexität der Modelle ohne SMOTE

Die Ergebnisse der Validierungsmetriken finden sich in Tabelle 3.

| **Validie-rung** | **Modell-art** | **AUC** (95%KI) | **BA** | **Accuracy** (95% KI) | **Sensitivität** (95%KI) | **Spezifität** (95%KI) | **F1-Score** (95%KI) |
|---|---|---|---|---|---|---|---|
| **In-sample** | logR | 0,77 (0,74/0,81) | 0,9 | 0,9 (0,89/0,91) | 0,09 (0,03/0,16) | 0,99 (0,98/1,0) | 0,15 (0,05/0,24) |
| | logR SMOTE | 0,79 (0,76/0,83) | 0,5 | 0,71 (0,68/0,74) | 0,69 (0,66/0,72) | 0,74 (0,67/0,81) | 0,71 (0,68/0,73) |
| | RFO | 0,7 (0,64/0,76) | 0,9 | 0,88 (0,86/0,9) | 0,11 (0,07/0,15) | 0,96 (0,95/0,98) | 0,16 (0,11/0,21) |
| | RFO SMOTE | 0,96 (0,91/1,0) | 0,5 | 0,89 (0,85/0,93) | 0,93 (0,81/1,03) | 0,86 (0,81/0,9) | 0,89 (0,84/0,95) |
| | GBM | 0,75 (0,72/0,79) | 0,9 | 0,89 (0,88/0,9) | 0,11 (0,04/0,17) | 0,98 (0,96/0,99) | 0,16 (0,06/0,25) |
| | GBM SMOTE | 0,95 (0,9/1,0) | 0,5 | 0,87 (0,8/0,95) | 0,85 (0,68/1,02) | 0,89 (0,84/0,94) | 0,86 (0,76/0,96) |
| **Out-of-sample** | logR | 0,81 | 0,91 | 0,91 | 0,14 | 0,99 | 0,23 |
| | logR SMOTE | 0,81 | 0,91 | 0,75 | 0,75 | 0,75 | 0,36 |
| | RFO | 0,7 | 0,91 | 0,9 | 0,16 | 0,97 | 0,22 |
| | RFO SMOTE | 0,64 | 0,91 | 0,47 | 0,72 | 0,45 | 0,2 |
| | GBM | 0,78 | 0,91 | 0,9 | 0,17 | 0,98 | 0,24 |
| | GBM SMOTE | 0,7 | 0,91 | 0,13 | 0,99 | 0,04 | 0,17 |

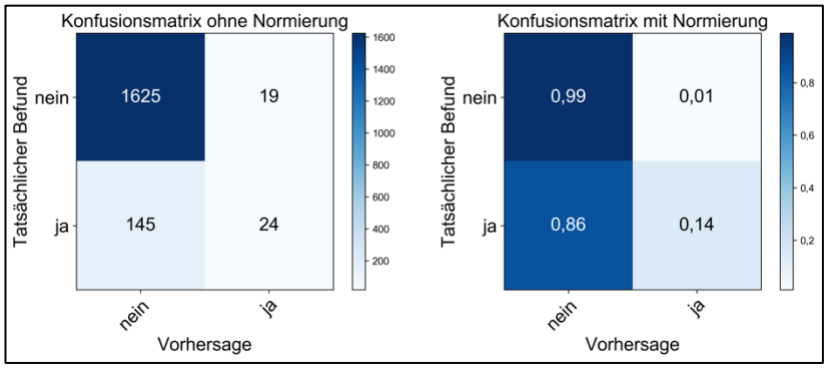
Tabelle 3: Metriken für die verschiedenen Validierungsstrategien: Area under the Curve (AUC), Baseline Accuracy (BA), Accuracy, Sensitivität, Spezifität und F1-Score. Ergebnisse der In-sample- und der Out- of-sample-Validierung.

Bei einem Schwellenwert von 0,5 haben alle Modelle in der In-sample-Validierung eine hohe Spezifität (zwischen 0,96 und 0,99) und eine niedrige Sensitivität (zwischen 0,09 und 0,11). Sie erreichen eine generell sehr hohe Accuracy zwischen 0,88 (KI: (0,86/0,9) und 0,9 (KI: (0,89/0,91). Ihre AUC liegt mit ihren 95% KI zwischen 0,64 und 0,81, wobei zu betonen ist, dass GBM und RFO in der In-sample-Validierung keine signifikant besseren Ergebnisse gegenüber der logR erzielen. In der Out-of- sample-Validierung erreicht die AUC Werte zwischen 0,7 und 0,81. Die Modelle kommen somit ebenfalls zu guten und im Fall der logR (AUC von 0,81) sogar zu sehr guten Ergebnissen.

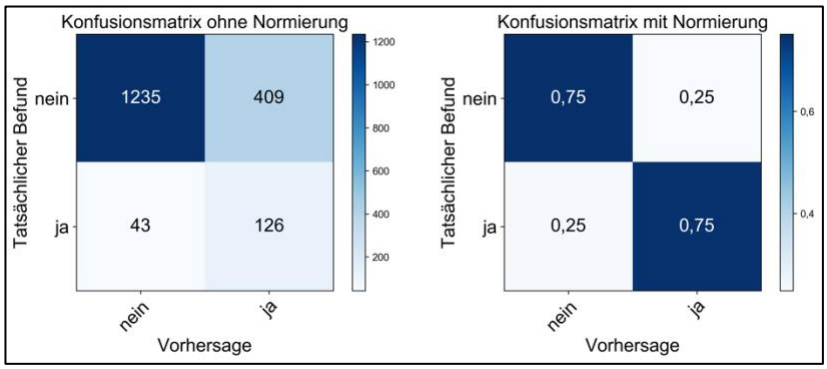
Demgegenüber stehen limitierte Werte für den F1-Score zwischen 0,22 und 0,24 in der externen Validierung, die sich unter anderem aus der niedrigen Sensitivität ergeben. Sieht man den F1-Score als Maß für die Genauigkeit und die Trefferquote der Modelle, so unterstreichen eine geringe TPR zwischen 0,14 und 0,17 und eine

hohe FNR zwischen 0,83 und 0,86 der Modelle die Ergebnisse (Abb.2a,c,e). Anhand des F1-Scores haben die komplexeren Modelle in der externen Validierung ebenfalls keinen Vorteil.

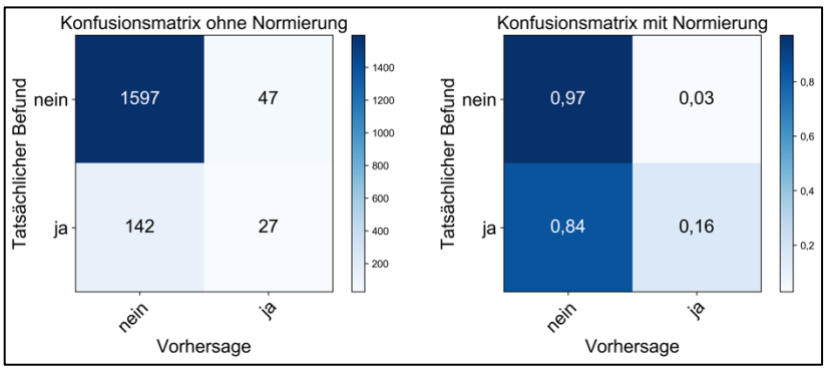
a



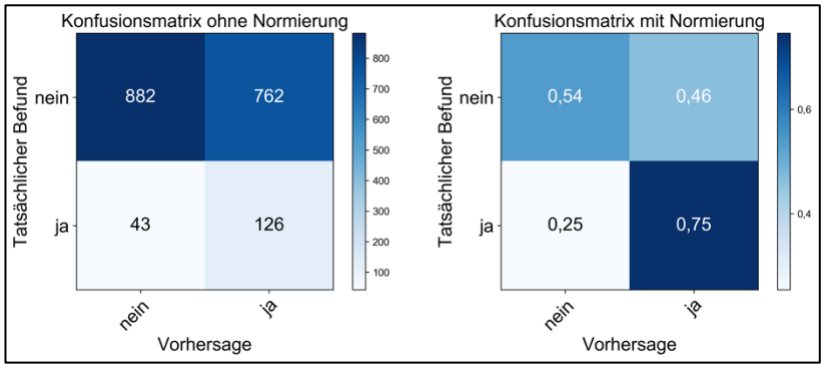

b



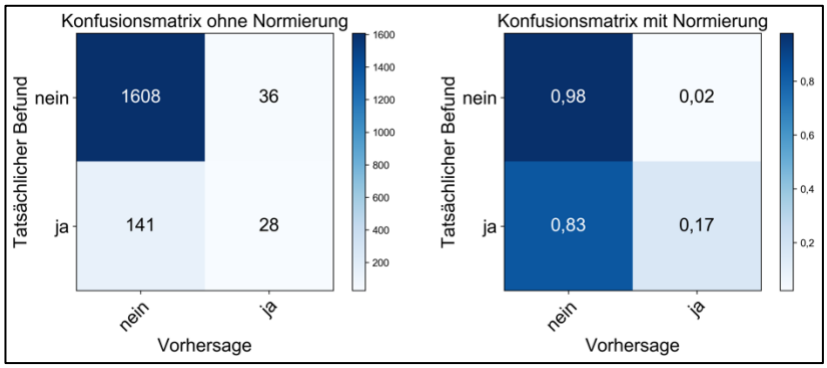

c




d



e




f

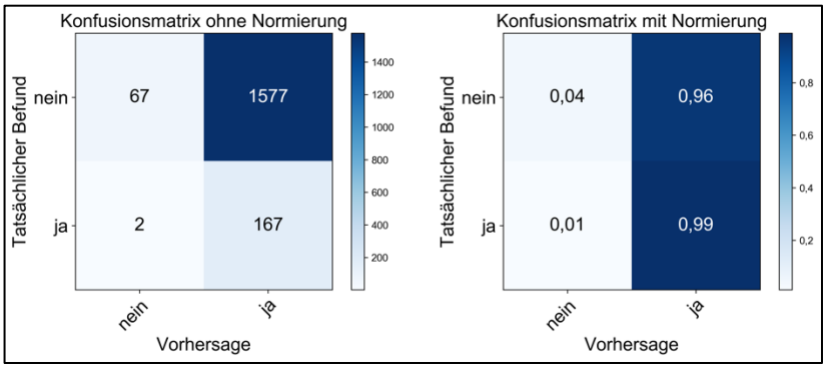



Abbildung 2: Konfusionsmatrizen der logistischen Regression (a,b), des Random Forests (c,d) und der Gradient Boosting Machine (e,f) ohne und unter der Anwendung der Synthetic Minority Over-sampling Technique (SMOTE).

## 5.3. Einfluss der Komplexität mit SMOTE

In der In-sample-Validierung sind die Werte für die Sensitivität mit SMOTE gegenüber der ohne SMOTE signifikant höher. Ihnen gegenüber stehen allerdings signifikant niedrigere Spezifitätswerte. So lässt sich eine Steigerung der Sensitivität von Mittelwerten zwischen 0,09 und 0,11 auf Werte zwischen 0,69 (logR) und 0,93 (RFO) erkennen, während die Spezifitätswerte zwischen 0,74 und 0,89 signifikant niedriger als die Werte mit SMOTE (zwischen 0,96 und 0,99) ausfallen. In Relation gesetzt, durch das Berechnen der AUC, erreicht der RFO eine AUC von 0,7 (KI:0,64/0,76) ohne und eine von 0,96 (KI:0,91/1,0) mit SMOTE. Die GBM steigert sich von 0,75 (KI:0,72/0,79) auf einen Wert von 0,95 (KI:0,9/1,0). Die logR liegt hier mit einer AUC von 0,79 (0,76/0,83) signifikant niedriger. Die ROC-Kurven mit und ohne SMOTE sind zur optischen Verdeutlichung in Abbildung 3 dargestellt.

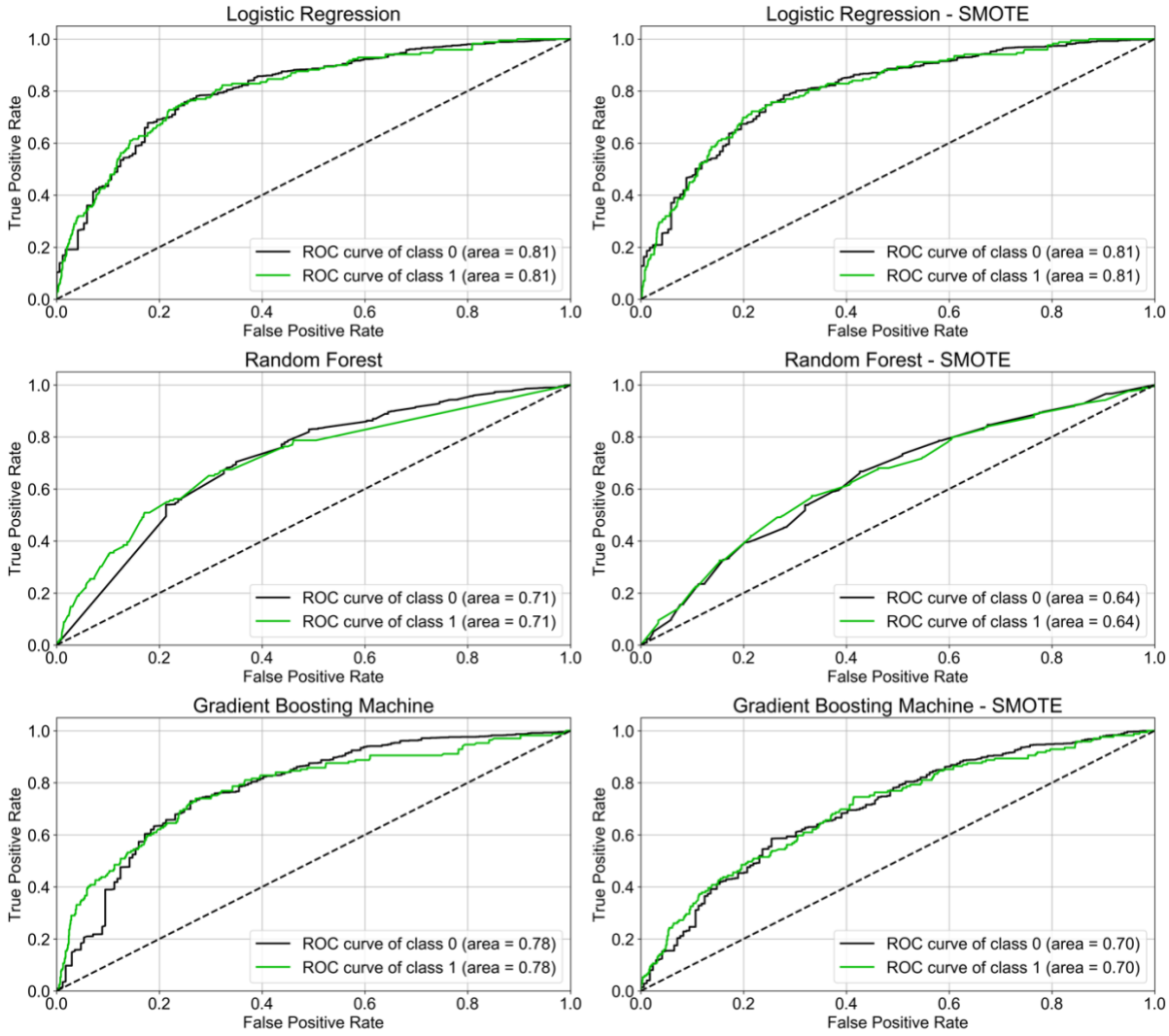



Abbildung 3: Receiver Operating Characteristic Curve (ROC Curve) der logistischen Regression, des Random Forests und der Gradient Boosting Machine mit und ohne SMOTE.

Demgegenüber zeigen in der Out-of-sample-Validierung, unter der Anwendung von SMOTE und bei einem Schwellenwert von 0,5 alle Modelle zwar ebenfalls eine erhöhte Sensitivität (zwischen 0,72 und 0,99), aber gerade die komplexeren Modelle eine viel niedrigere Spezifität (0,04 und 0,45) und auch die logR eine geringere, dennoch moderate Spezifität von 0,75 auf. Diese Veränderung ist bei der GBM besonders ausgeprägt. Die Sensitivität beträgt 0,99 und die Spezifität lediglich 0,04. In Relation zu diesen Werten fällt die Accuracy der komplexen Modelle mit SMOTE von 0,9 auf moderate 0,47 (RFO) und von 0,9 auf limitierte 0,13 (GBM) und bei der logR von 0,91 auf 0,75. Die AUC der komplexeren Modelle ist entsprechend der Veränderungen von Spezifität und Sensitivität unter SMOTE ebenfalls niedriger (bei RFO von 0,70 auf 0,64 und GBM von 0,78 auf 0,7). Die AUC der logR bleibt mit 0,81 konstant. Hierbei sind

die unterschiedlichen Reaktionen der Modelle auf das Over-sampling, das Verhältnis indem sich Spezifität und Sensitivität verändern, zu beachten (Abb.4).

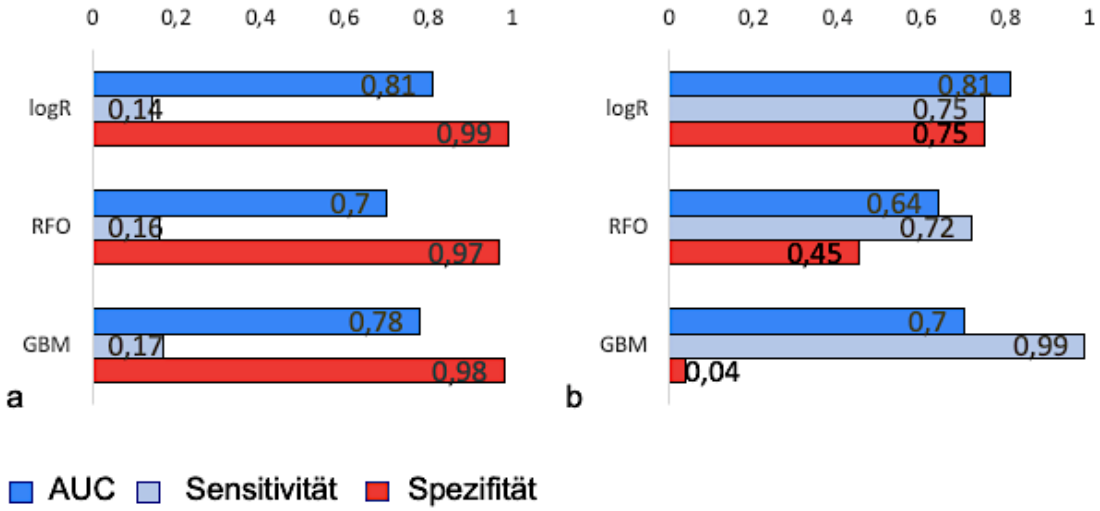



Abbildung 4: (a) ohne SMOTE und (b) mit SMOTE (Synthetic Minority Over-sampling Technique): Balkendiagramme der AUC (Area under the Curve), der Sensitivität und der Spezifität (x-Achse) der logistischen Regression (logR), des Random Forests (RFO) und der Gradient Boosting Machine (GBM) auf der y-Achse.

Die komplexeren Modelle weisen unter der Verwendung von SMOTE in der Out-of-sample-Validierung also keine höheren AUC-Werte als die weniger komplexe logR auf. Tatsächlich steigt die FPR aller Modelle von Werten zwischen 0,01 und 0,03 auf Werte zwischen 0,25 (logR) und 0,96 (GBM). Modellspezifisch heißt das für die logR einen Anstieg der falsch positiven Vorhersagen von 19 auf 409, für RFO von 47 auf 762 und bei der GBM von 36 sogar auf 1577 (Abb. 2). Hier kommt zum Tragen, dass eine Erhöhung der tatsächlich positiven Ergebnisse zu einer Erhöhung der falsch positiven Ergebnisse (FPR, 1-Spezifität) führen kann und dies eine Abnahme der Precision zur Folge hat (Chawla, 2005). Diese beeinflusst den F1-Score, welcher mit 0,17 (GBM) und 0,20 (RFO) für die komplexen Modelle erstens niedriger als der F1-Score ohne SMOTE und zweitens niedriger als der der logR mit 0,36 ist.

## 5.4. Vergleich der Ergebnisse der internen mit denen der externen Validierung

Allgemein lässt sich nur unter der Verwendung von SMOTE die oftmals eintretende Verschlechterung der Ergebnisse von der In-sample- Validierung gegenüber der Out-of-sample Validierung (Bouwmeester et al., 2012, Sun et al., 2011) und dies auch nur

bei den komplexeren Modellen erkennen. Ohne SMOTE findet kein Überschätzen statt. So nehmen die Werte für die Validierungsmetriken in der Out-of-sample-Validierung sogar leicht zu (z.B. AUC logR von 0,77 auf 0,81) oder bleiben konstant (AUC RFO intern und extern 0,7). Vergleicht man die Ergebnisse der internen mit denen der externen Validierung unter Verwendung von SMOTE, so fallen die extrem hohen Werte der AUC (zwischen 0,95 und 0,96) und des F1-Scores (zwischen 0,86 und 0,89) in der In-sample Validierung der komplexen Modelle sowie der In-sample-F1-Score der logR (0,71) auf. Sie stehen Out-of-sample Ergebnissen mit AUC-Werten von 0,64 (RFO) und 0,7 (GBM) und F1-Scores von 0,36 (logR), 0,20 (RFO) und 0,17 (GBM) gegenüber. Wie bereits beschrieben, steigen und fallen die Sensitivität und die Spezifität der Modelle nicht im gleichen Maße. Für die GBM und den RFO führt dies zu einem starken Abfall der Accuracy (0,13 (GBM) und 0,47 (RFO)).

Allgemein lässt sich nur unter der Verwendung von SMOTE ein signifikanter Unterschied zwischen den In-sample- und den Out-of-sample-Ergebnissen der AUC der komplexeren Modelle, also von RFO und GBM feststellen. Die In-sample-Ergebnisse sind hier signifikant höher. Ohne SMOTE gibt es keinen signifikanten Unterschied.

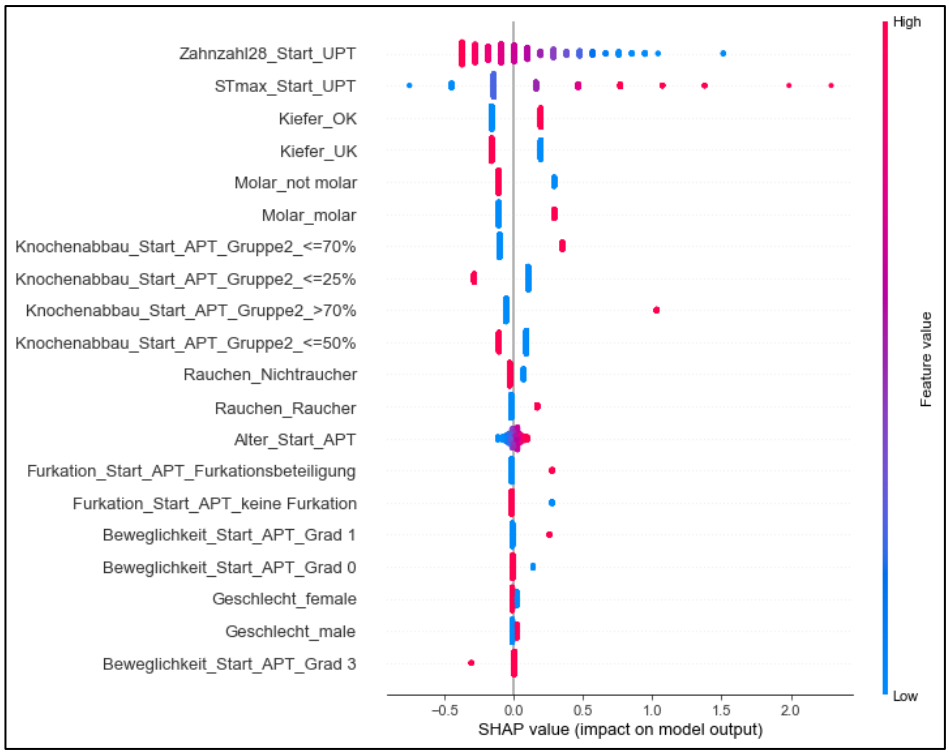
## 5.5. SHAP-Wert als Interpretationshilfe

### 5.5.1. SHAP Summary Plot

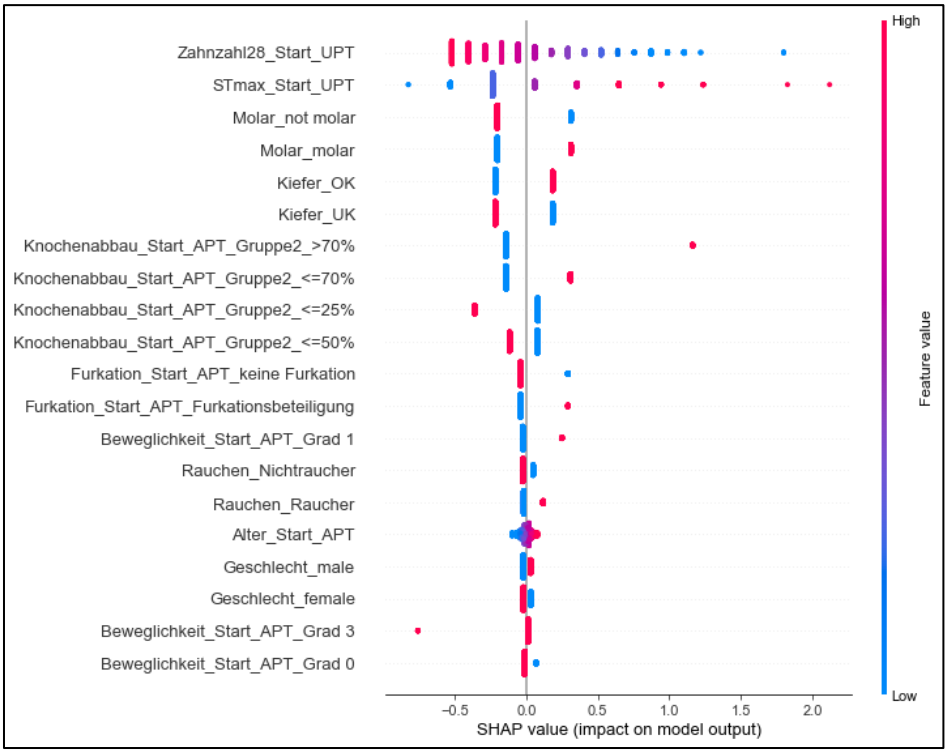
Anhand der Summary Plots von SHAP kann man erkennen, dass die drei Modelle auf globaler Ebene Variablen unterschiedlich stark für ihre Prognosen heranziehen. Die Modelle werten ohne SMOTE beispielsweise alle die Anzahl an Zähnen zu Beginn der UPT und die maximalen Sondierungstiefen als Prädiktoren mit dem größten Einfluss auf Rang eins und zwei (die Prädiktoren sind in der Grafik ihrer Relevanz nach absteigend sortiert). Zur weiteren Erklärung der Summary Plots: Die Lage auf der x-Achse zeigt an, ob und wie stark die Variabel das Ergebnis positiv oder negativ beeinflusst (SHAP Value) und die Farbe die Ausprägung des Merkmals (rot: hoch oder ja; blau: niedrig oder nein). Das Modell wird im Ganzen, also unter Bezugnahme aller Beobachtungen (global) erklärt, wobei man auch den Einfluss jeder einzelnen Beobachtung ablesen kann. Für die logR bedeutet dies beispielsweise, dass die Zahnanzahl auf Rang eins das Ergebnis bei hohen Werten (hoher Restzahnbestand, rot) negativ, sprich in Richtung Zahnerhalt, beeinflusst. Auf die maximale

Sondierungstiefe folgen der Kiefertyp (Oberkiefer (OK) oder Unterkiefer (UK)) auf Rang drei und vier und auf Rang fünf und sechs die Zahnart (Nicht-Molar oder Molar) (Abb. 5a). Unter der Verwendung von SMOTE zieht das gleiche Modell hingegen den Typ des Zahns stärker (Platz drei und vier in der Rangfolge) als die Kieferart (Abb. 5b) heran. Das Alter spielt für die logR eine untergeordnete Rolle (Rang 13 ohne und Rang 16 mit SMOTE).

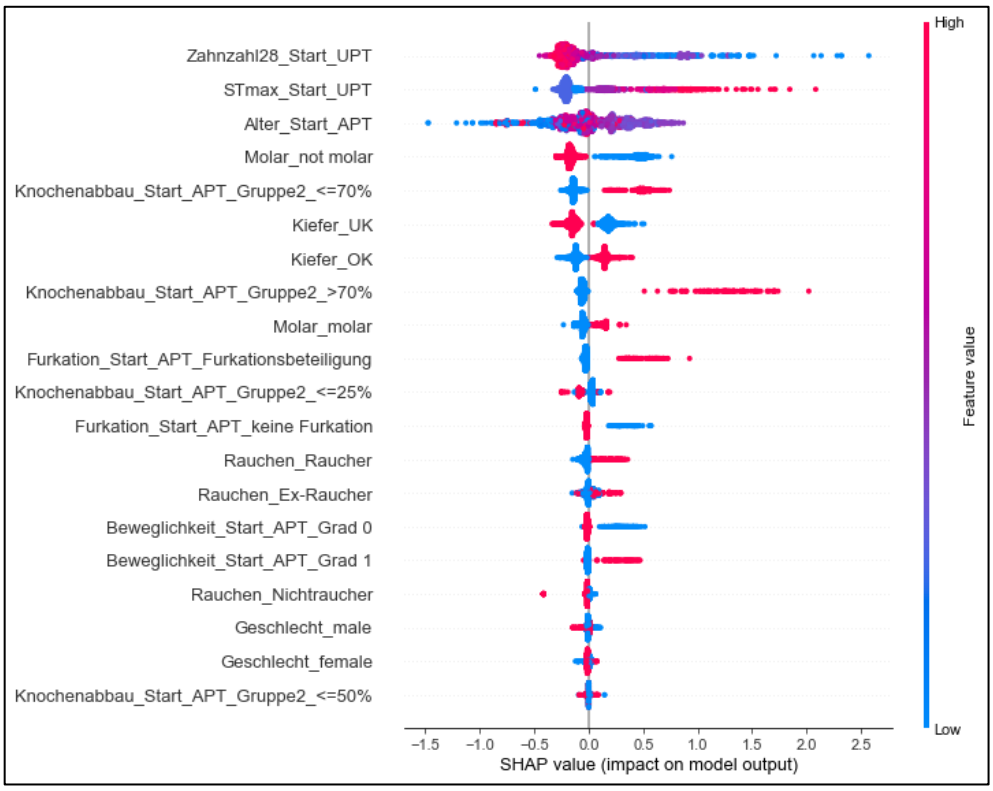
a




b



c




d

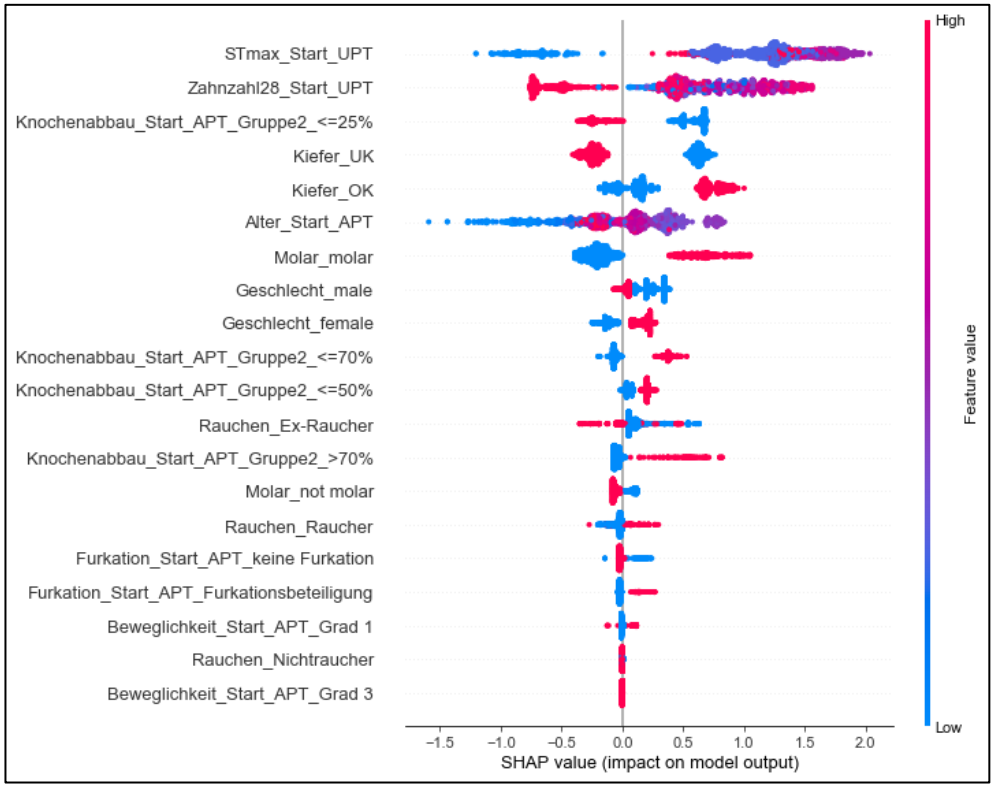
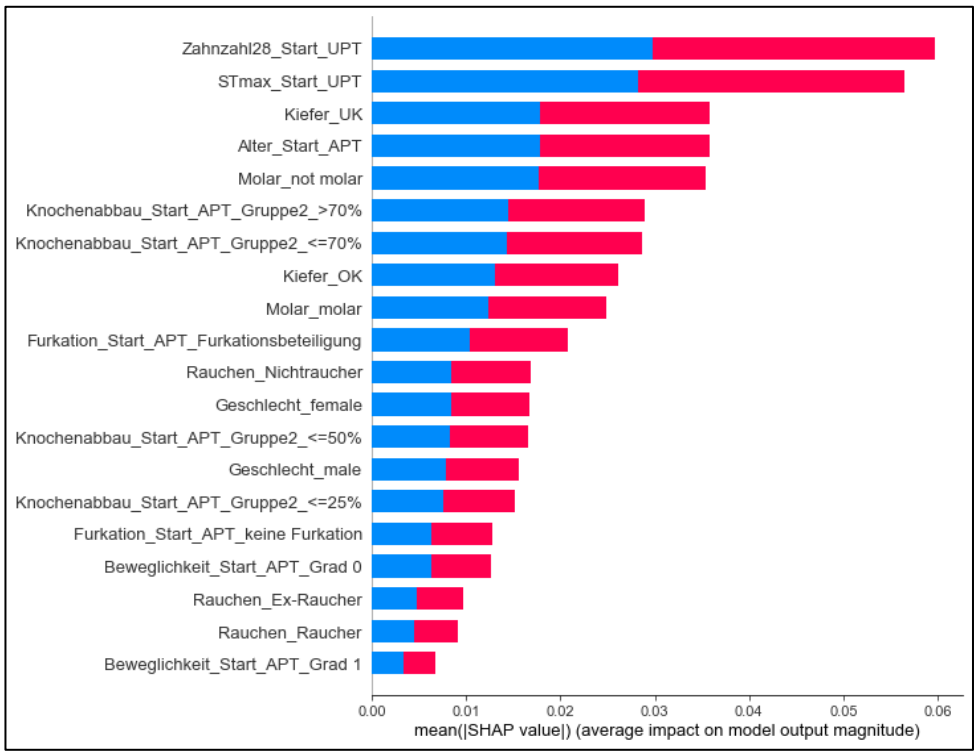
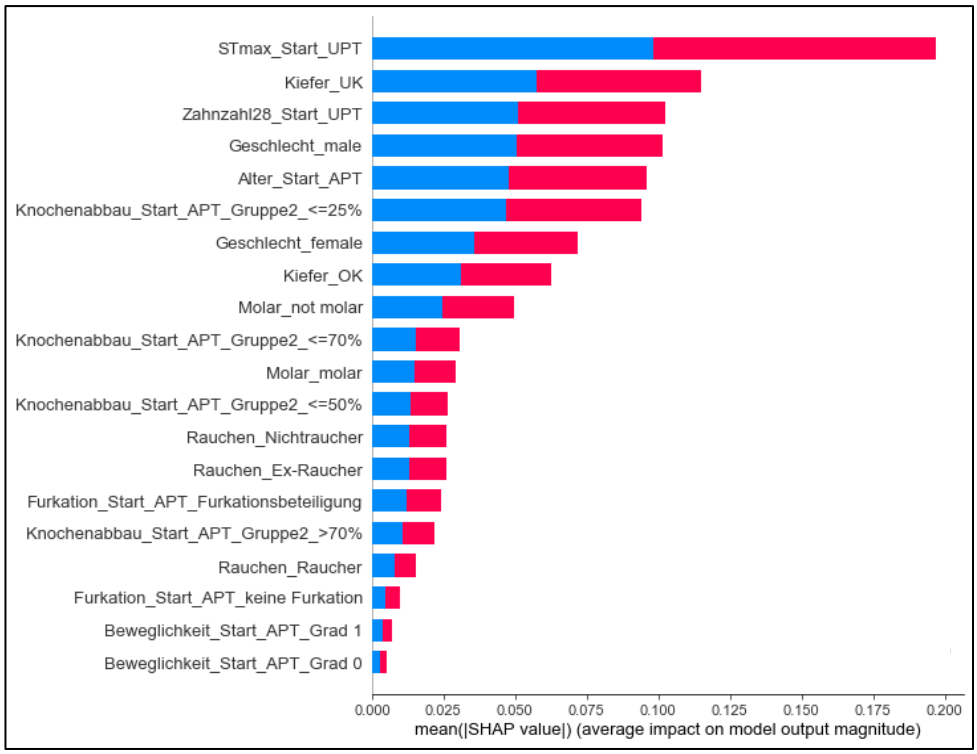



Abbildung 5: Summary Plot der logistischen Regression *und der Gradient Boosting Machine o*hne Synthetic Minority Over-sampling Technique *(a,c) und mit (b,d):* x-Achse: SHAP Value (-1 bis 2,5); y-Achse: Variablen sortiert nach ihrer Relevanz; Einzelne Beobachtungen unterschiedlich groß dargestellt: je größer, desto relevanter die Variable; rot: Ausprägung hoch oder ja; blau: Ausprägung niedrig oder nein.

a




b



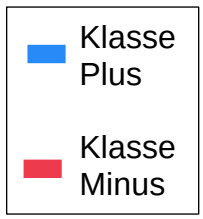




Abbildung 6: Vereinfachter Summary Plot des Random Forests ohne (a) und mit Synthetic Minority Over-sampling Technique (b): x-Achse: Mittelwert der SHAP Values der Variablen; y-Achse: Variablen sortiert nach ihrer Relevanz für die Vorhersage; blau: Klasse Plus (Zahnverlust); rot: Klasse Minus (Zahnretention).

Um den Unterschied zu den anderen Modellen aufzuzeigen, ist es beim RFO nach der Zahnanzahl und den Sondierungstiefen von nächst höherer Bedeutung, ob es sich um einen Zahn im UK (Rang drei) handelt und wie alt der Patient (Rang vier) ist (Abb. 6a). Der Summary Plot ist für den RFO hier vereinfacht dargestellt. Er zeigt im Gegensatz zu der Abbildung 5 nur den Mittelwert des Einflusses der Variablen auf die Vorhersagen im Ganzen ohne die Ausprägungen der einzelnen Beobachtungen darzustellen. Das Alter spielt auch für die Vorhersage der Gradient Boosting Machine eine große Rolle (Rang drei, Abb.5c). Die Ausprägungen beeinflussen die Vorhersage vielfältiger, als die der logR (Vgl.: Abb. 5a und b). So wird die Vorhersage von Zahnverlust bei der logR eindeutig durch hohes (rote Markierung) Alter des Patienten positiv beeinflusst, während in Abbildung 5c (GBM) eine mittleres (violette Markierung) Alter die Vorhersage stark positiv beeinflusst. Hohes Alter eines Patienten hat sogar einen leicht negativen Einfluss. Allgemein erscheint der relative röntgenologische Knochenabbau mit unterschiedlichen Ausprägungen noch auf höheren Rängen, während die Furkationsbeteiligung, der Raucherstatus, die Beweglichkeit des Zahns und das Geschlecht des Patienten weniger relevant für die Modelle zu sein scheint (Abb. 5 und 6). Hier handelt es sich nur um Tendenzen, da es große, modellspezifische Unterschiede insbesondere unter der Verwendung von SMOTE bei komplexeren Modellen (Abb.5d und 6b) gibt.

### 5.5.2. SHAP Force Plot

Anhand der SHAP Force Plots kann man den SHAP Value der Variablen auf lokaler Ebene für einzelne Beobachtungen und die Richtung in welche jede Variable das lokale Ergebnis, die Wahrscheinlichkeit der Vorhersage von Zahnverlust (output value, zu Deutsch: Ausgabewert), vom Mittelwert der Vorhersagewerte anhand der Trainingsdaten (base value) weg beeinflusst. Bei dem aufgeführten Beispiel handelt es sich um eine tatsächlich positive, also korrekt vorhergesagte Beobachtung (tatsächlicher Zahnverlust) der logR. In der abgebildeten Beobachtung (Abb. 7) hat das Merkmal Knochenabbau mit einer Ausprägung von >70% den größten positiven (längster roter Pfeil nach rechts) Einfluss auf die Prognose Zahnverlust (SHAP Value: +4,253). Die Rangfolge bei dieser Beobachtung entspricht also nicht der des ganzen Modells. Es folgt die Zahnanzahl mit einem negativen SHAP Value von -2,493, da der Patient lediglich 15 Zähne zu Beginn der UPT hatte. Auch dies begünstigt die Vorhersage „Zahnverlust“. Der SHAP Value variiert bei nicht dichotomen Variablen je

nach Ausprägung. Einen leicht negativen Einfluss auf die Prognose Zahnverlust (kurzer blauer Pfeil nach links) hat die Tatsache, dass der Zahn sich im Unterkiefer befindet. Das Merkmal „Oberkiefer“ hat eine negative Ausprägung (-0,9118). Insgesamt überwiegen die positiven nach rechts (positiver Bereich) orientierten Einflüsse auf den base value  und die Wahrscheinlichkeit eines Zahnverlustes beträgt 0,12 (output value). Dieser Wert liegt weit über dem base value von -2,664. Die Prognose lautet Zahnverlust und in diesem Fall stimmt die Vorhersage mit dem Befund des Patienten überein.

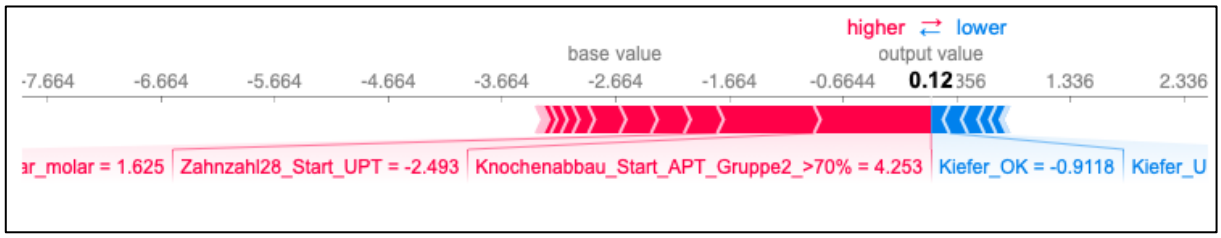



Abbildung 7: Ausschnitt eines Force Plots einer Beobachtung (eines Zahns), die korrekt positiv klassifiziert wurde. Pfeilrichtungen: Richtung in die der durchschnittliche Vorhersagewert anhand des Trainingsdatensatzes (base value) durch die Variablen beeinflusst wird, um den Vorhersagewert, die Wahrscheinlichkeit dieser Beobachtung, dass Zahnverlust prognostiziert wird (output value) zu erhalten. Rot: erhöht die Wahrscheinlichkeit; Blau: senkt die Wahrscheinlichkeit.

# 6. Diskussion

Die Verwendung von Vorhersagemodellen in der Zahnmedizin soll neben dem gesundheitlichen Nutzen für den Patienten auch seine orale Lebensqualität steigern (Bäumer et al., 2018, El Sayed et al., 2019) und zukünftige Kosten im Gesundheitswesen und für den Patienten selbst besser abschätzbar machen. Unerwarteter Zahnverlust und damit verbundene Risiken, eine Minderung der Behandlungseffizienz sowie subjektive oder nicht personalisierte Entscheidungen zur Zahnextraktion sollen verhindert werden. Um diesen Ansprüchen gerecht zu werden, müssen Modelle rigoros entwickelt werden und zuverlässig und generalisierbar sein.

## 6.1. Interpretation

Die verschiedenen erhobenen Metriken in dieser Arbeit haben unterschiedlich hohen analytischen Wert. So ist die Accuracy mit ihren sehr hohen Werten für die unausgeglichenen Daten (ohne SMOTE) keine geeignete Metrik, um die Genauigkeit der Modelle darzustellen (Blagus and Lusa, 2015, Chawla, 2005, Sun et al., 2011). Sie

zeigt im Vergleich zu der Baseline Accuracy dennoch, dass die Modifikationen der Algorithmen zu keinen Vorteilen geführt haben. Das Modell oder die Daten also ungeeignet sind. Unter der Verwendung von SMOTE spiegelt die Accuracy die Leistungen der Modelle wesentlich realistischer wider. So klassifiziert die logR im Zuge der externen Validierung nur noch moderate 75% der Beobachtungen richtig und die GBM sogar nur 13% aller Beobachtungen. Andere Metriken wie die AUC beschreiben die Performance der Modelle unabhängig von der Klassenverteilung und sind mit und ohne SMOTE aussagekräftig. Während sich gute bis sehr gute AUC-Werte für die eigentlichen Daten ergaben, wurden schlechtere Ergebnisse unter der Anwendung von SMOTE berechnet. Zwar erhöht sich sinngemäß die Sensitivität (Zahnverlust wird korrekt vorhergesagt), die Spezifität (Zahnretention korrekt vorhergesagt) fällt aber gerade bei den komplexen Modellen gravierend ab. Im Verhältnis ist der negative Einfluss auf die Spezifität bei komplexeren Modellen größer als der positive Effekt auf die Sensitivität, die AUC fällt (Abb. 4).

Die AUC als alleinstehende Metrik zu betrachten, ohne der Sensitivität und der Spezifität Aufmerksamkeit zu schenken, ist für die Interpretation der Ergebnisse im medizinischen Sinne nicht auszureichend. Sie ermöglicht aber einen Vergleich von Modellen und präsentiert darüber hinaus ihre Stärke in ihrer Unabhängigkeit gegenüber den Klassenproportionen. So zeigt ein sehr guter Wert in der Out-of-sample-Validierung von 0,81 mit und ohne SMOTE, dass die logR einerseits genauso gut mit den unausgeglichenen Daten wie mit einer Over-sampling-Technik wie SMOTE umgehen kann und andererseits, dass sie den komplexeren Modellen in beiden Fällen überlegen ist. Die logR reagiert sehr robust auf die Veränderungen unter anderem der Häufigkeitsverteilung der Klassen (Perlich et al., 2003). Ein gutes Maß zur Ergänzung ist der F1-Score. Er vereint die Fähigkeit des Modells Zahnverlust richtig zu prognostizieren mit der allgemeinen Trefferquote. Erwartungsgemäß erhalten die Modelle niedrige Werte anhand der unausgeglichenen Daten, da die Klassifikatoren kaum Ereignisse zum Erlernen von Mustern hatten. Die ebenfalls limitierten Ergebnisse mit SMOTE lassen sich anhand der allgemeinen Trefferquote erklären. Nach der Anwendung von SMOTE steigt zwar die Sensitivität der Modelle, es verschlechtern sich aber die Werte der Precision (s.o.), da zunehmend falsch positive Prognosen gestellt werden (Zahnverlust wird vorhergesagt, tritt aber nicht ein). Die sehr hohen Werte für die AUC und den F1-Score der komplexen Modelle in der In-sample-Validierung unter Verwendung von SMOTE führen zu einer Überschätzung

ihrer Diskriminationsfähigkeit. Dies liegt daran, dass es zum Overfitting kommt (Ochani et al., 2019) und wird durch das Stattfinden der Kreuzvalidierung vor dem Over-sampling noch verstärkt (Blagus and Lusa, 2015). Die neu generierten, künstlichen Beobachtungen ähneln denen der Minoritätsklasse (Beobachtungen mit eintretendem Zahnverlust) sehr stark. Werden diese vor der Kreuzvalidierung synthetisiert, so kann es sein, dass Beobachtungen und nach ihrem Vorbild generierte neue Beobachtungen miteinander validiert werden und füreinander unrealistisch akkurate Prognosen stellen können. Die Beobachtungen im Validierungsdatensatz können aber stark von diesen variieren, folglich nicht richtig klassifiziert werden. Ein Phänomen, das gerade bei den komplexeren Algorithmen auftritt, da die Modelle sich zu gut an die Entwicklungsdaten anpassen (Krois et al., 2019, Sun et al., 2011). Im Testdatensatz gibt es keine künstlichen Daten. Die Modelle klassifizieren die positiven Beobachtungen (Zahnausfälle) weitgehend richtig, tun dies aber weniger spezifisch, so dass teils negative Beobachtungen (Zahnretentionen) ebenfalls als positiv vorhergesagt werden. Neben der Überschätzung findet also auch noch eine Übergeneralisierung (Yen and Lee, 2006) nach der Applikation von SMOTE statt. Diese entsteht laut Yen aus der Missachtung der negativen Beobachtungen. Die Algorithmen erwarten in größerem Maße positive Ereignisse, da die Entscheidungsräume für diese gewachsen sind. Dies spiegelt sich in der gesteigerten FPR wider. Demgegenüber konnten Entwicklungsdaten für negative Ereignisse noch von den Algorithmen „auswendig“ gelernt werden, während sie an unbekannten Daten bezüglich der negativen Prognosen an Genauigkeit verlieren. Die Spezifität und die Accuracy fallen in der Out-of-sample-Validierung dementsprechend stark ab. Dies kann daran liegen, dass die positiven Ereignisse ehemals extrem selten waren und die Gefahr einer Durchmischung (Branco et al., 2016) der Entscheidungsräume somit erhöht ist (Yen and Lee, 2006). Verstärkt wird dies durch die Vergrößerung der Regionen für die positive Klasse. Neben den natürlichen, vorher eventuell als FN prognostizierten Beobachtungen, liegen nun auch viele künstlich generierte im Bereich der ehemaligen Majoritätsklasse (negative Klasse, Zahnretention). Die Daten sind nicht mehr klar zu separieren. Negative Beobachtungen, die in die k-Nearest-Neighborhood von positiven Beobachtungen und damit nach dem Over-sampling sehr wahrscheinlich in einen Entscheidungsraum für diese fallen, werden falsch positiv prognostiziert. Im Validierungsdatensatz bleiben die Entscheidungsräume bestehen, es gibt aber keine künstlichen Beobachtungen mehr. Die falsch positiven Prognosen nehmen einen

größeren prozentualen Anteil an. Folglich sinken die Accuracy und die Spezifität, wie bereits erwähnt. Die logR gilt bei schlecht zu separierenden Daten den störanfälligeren, komplexeren Modellen als überlegen (Perlich et al., 2003), was sich anhand der vorliegnden Ergebnisse bestätigen lässt. Gründe könnten ihre Robustheit gegenüber einer moderaten Menge an Rauschen und ihre geringe Varianz sein (Perlich et al., 2003), die auch die geradezu unveränderte Leistung mit und ohne SMOTE erklären. Die logR ist relativ rigide und kann sich daher nicht zu gut an Daten anpassen. Ihre Entscheidungsräume werden dadurch weniger stark durchmischt. Komplexere Algorithmen sind hingegen sehr sensitiv gegenüber ihrer Trainingsdaten und weisen eine hohe Varianz auf (Perlich et al., 2003). Allgemein ist also festzustellen, dass Modelle unter SMOTE eher positive Ergebnisse vorhersagen. Außerdem führt die Zunahme der Sensitivität zu einem Abfall der FNR. Geht man davon aus, dass die falsche Vorhersage von nicht eintretendem Zahnverlust, weniger schlimm als ein unerwarteter, nicht prognostizierter Zahnverlust ist, so wäre diese Veränderung im Grunde genommen als positiv zu werten. Allerdings steigt die FPR in den Fällen der komplexen Modelle in nicht tragbarem Maße, so treten 55% (RFO) oder gar 96% (GBM) der vorhergesagten Zahnverluste gar nicht ein. Für die logR nehmen die fehlerhaften Prognosen jeweils 25% der positiven und negativen Vorhersagen an. Im Vergleich zu den 86% (n=145 Zähne), welche ohne SMOTE falsch negativ eingeschätzt werden würden, ist dies eine Verbesserung. Sie führt aber nicht zu einer klinischen Anwendbarkeit der logR mit SMOTE, da eine FPR von 25% bedeutet, dass 409 Zähne falsch positiv prognostiziert werden. Eine FNR von 25% bedeutet, dass 43 Zähne laut Modell erhalten bleiben, tatsächlich aber Zahnverlust eintritt. In das Modell vertrauende Behandler könnten falsche, auch kostenintensive Maßnahmen zum Erhalt oder zur Restauration ergreifen (Schwendicke et al., 2017, Schwendicke et al., 2014). Die Modelle unter der Verwendung von SMOTE sind klinisch also nicht anwendbar und komplexe Modelle sind darüber hinaus auch nicht generalisierbar, da sie durch ihre Übergeneralisierung in der Entwicklungsphase an Diskriminationsstärke verlieren. Es wird deutlich, dass SMOTE für das Problem der unausgeglichenen Daten von Prognosemodellen zum Einschätzen von zukünftigen Zahnverlust in complainten Patientenkohorten während der parodontalen Nachsorgephase, keine geeignete Lösung darstellt.

An diesem Punkt ist zu erwähnen, dass Chawla et al. (2002) die Kombination aus SMOTE und Under-sampling empfehlen, was in der Anwendung von SMOTE für die

Modelle dieser Studie nicht umgesetzt wurde. So sollte zunächst ein Over-sampling bis zu einem von dem Stichprobenumfang abhängigen Wert und im Anschluss ein Under-sampling durchgeführt werden. Um diesen Wert zu erhalten, sollten verschiedene Raten an Over-sampling (100%, 200%, usw.) vorgenommen und über ROC-Kurven in Kombination mit Under-sampling verglichen werden.

Ohne SMOTE erzielen die Modelle bessere oder zumindest vergleichbare Ergebnisse in ihrer AUC. Es wird aber deutlich, dass eine Sensitivität, sprich eine Vorhersage von tatsächlichem Zahnverlust, in nur 9-11% Prozent der Fälle nicht ausreicht. Somit ist zu klären, wo sich der Ursprung des Problems befindet und welche Möglichkeiten bestehen, dieses zu lösen.

## 6.2. Implikationen

Neben dem offensichtlichen Missverhältnis zwischen Klassen wurde diskutiert, ob die Ursache für eine schlechtere Performance von Modellen ihren Ursprung auch in der Verteilung innerhalb der Klassen selber haben könnte. So könnte die Überschneidung von Entscheidungsräumen (Overlapping), eine zu geringe Dichte der Trainingsdaten, verzerrte Daten in unausgeglichenen Räumen, die schwierige Identifikation von Regionen mit kleinen Disjunktionen oder die unterschiedlichen Wahrscheinlichkeitsverteilungen in Trainings- und Validierungsdaten der Grund für die geringe Sensitivität sein (Branco et al., 2016). Dies gilt es für die Daten zum Zahnverlust auszuschließen oder in weitere Überlegungen zum Umgang mit den Daten mit einzubeziehen.

Andere Autorengruppen haben die Problematik der Klassenunausgeglichenheit aufgegriffen und Ansätze wie Under-sampling Techniken beschrieben, welche zu besseren Ergebnissen unter anderem als SMOTE (Blagus and Lusa, 2015) in den meisten Datensets geführt haben sollen. Obwohl es hier auch Studien mit entgegengesetzten Ergebnissen gibt (Naseriparsa and Kashani, 2013). Bei der generellen Anwendung von Sampling-Methoden ist die richtige Anwendung in Kombination mit der internen Validierung, explizit der Kreuzvalidierung zu beachten (Blagus and Lusa, 2015). Und zwar sollten die Under- und Over-sampling Techniken in bereits geteilten Daten nur auf die $k - 1$ -Teilmengen appliziert werden, die zum

Trainieren des Klassifikators verwendet werden, da es sonst beim Over-sampling zu optimistischen Ergebnissen in der In-sample-Validierung kommt, die in der Out-of-sample Validierung nicht erreicht werden können, da die Modelle nicht generalisierbar sind, wie bereits ausführlich diskutiert.

Außerdem sollten allgemein bereits vorhandenen Modelle erweitert und verbessert werden, anstatt sie bei nicht ausreichender Performance einfach zu verwerfen (Collins et al., 2016, Cowley et al., 2019). Hierfür scheinen sich Modelle mit Verwendung einer logR zu eignen, da sie Machine Learning Algorithmen in nichts nachstehen sollen (Christodoulou et al., 2019) und entsprechend der Ergebnisse dieser Studie (mit und ohne Sampling-Ansatz) im Umgang mit den Daten zum Zahnverlust zu bevorzugen sind. So wird zu einer Identifikation, der am besten geeigneten Algorithmen für Problemstellungen aufgerufen (Christodoulou et al., 2019), um die Entwicklung eines anwendbaren Modells voranzutreiben. Es sollte auf jeden Fall eine Analyse über den klinischen Nutzen und Konsequenzen erfolgen, um beispielsweise einschätzen zu können, welche Vorhersage für den Behandler, aber auch individuell für den Patienten als schlimmer zu werten ist (Vickers and Elkin, 2006). Außerdem sollte für ein Modell, welches gute Ergebnisse in mehreren „Broad“- Validierungen erzielte, mittels einer umfangreichen Einfluss-Studie („impact study“, (Cowley et al., 2019, Toll et al., 2008), der tatsächliche Einfluss nicht nur auf die Behandlung, den Behandler und den Patienten, sondern auch auf den klinischen Alltag und die Kosten bestimmt werden. Sollten die vielversprechendsten Ansätze keine Verbesserungen erzeugen oder externe („Broad“-) Validierungen diese nicht bestätigen, so ist davon auszugehen, dass man anhand der zur Verfügung stehenden Daten nicht zu genaueren Prognosen kommen kann. So könnten zum Beispiel die Prädiktoren nicht aussagekräftig genug sein (Cowley et al., 2019). Wie in einem systematischen Übersichtsartikel aus dem Jahre 2018 bereits vorgeschlagen, könnten dann neue Prädiktoren wie Sulkusflüssigkeit, Biomarker im Speichel oder mikrobiologische Informationen in Erwägung gezogen werden (Du et al., 2018). Deren Einfluss auf die Destruktion der parodontalen Erkrankung ist aber immer noch unzureichend untersucht beziehungsweise wird deren Validität kontrovers diskutiert (AlMoharib et al., 2014, AlRowis et al., 2014, Gursoy et al., 2011, Schmalz et al., 2016). Neuere Studien befassen sich mit dem Zusammenhang zwischen Zahnverlust und Compliance der Patienten (Petit et al., 2019) oder bestätigen die Zusammenhänge zwischen der Menge an Zigaretten pro Tag und der Ausprägung von Diabetes (gemessen am

HbA1c-Wert) mittels „grading“, also der Progressions-Rate der Parodontitiserkrankung, wie es in der neuen Klassifikation von (Tonetti et al., 2018) beschrieben wurde. Die Ergebnisse zeigen eine maßabhängige Vorhersagekraft bezüglich des Eintretens von Zahnverlust (Petit et al., 2019). So sollte man auch das Ausmaß der Variablen in Anbetracht ziehen. Das Geschlecht eines Patienten hingegen zeigt in einer aktuellen Studie (Ravida et al., 2019) keinen signifikanten Zusammenhang mit der Anzahl von Zahnverlusten aus parodontalen Gründen und auch in dieser Studie lässt sich anhand der Variablenrangfolge der SHAP Summary Plots darstellen, wie wenig Einfluss diese Variable auf die Vorhersagen der Modelle hat. Hier sind modellspezifisch Unterschiede zu erkennen, welche insbesondere bei den komplexeren Modellen unter der Verwendung von SMOTE auftreten, welche aufgrund ihrer Unzuverlässigkeit allerdings nicht repräsentativ sein sollten. Eine große Menge von Variablen erhöht die Gefahr für Overfitting und Rauschen (Ali, 2015). Folglich könnte man das Geschlecht als Prädiktor exkludieren, wenn es nicht zum Beispiel im Zusammenhang mit der Compliance eine Rolle spielt (Petit et al., 2019).

Eine systematische, rigorose Herangehensweise (Cowley et al., 2019, Du et al., 2018) wie sie hier zur weiteren Validierung und zum Aktualisieren von Modellen angedeutet und untersucht wurde, ist von zentraler Bedeutung, um eine Stagnation in der Entwicklung eines klinisch anwendbaren, generalisierbaren Modells zu verhindern. Um die Forschung voranzutreiben, müssen bereits bestehende Richtlinien, Erkenntnisse und Reviews beachtet werden.

## 6.4. SHAP

Im Rahmen dieser Studie wurde eine Interpretationshilfe für die Ergebnisse der Vorhersagemodelle vorgestellt. Anhand der Force Plots kann die individuelle Relevanz (Shapley Values, SHAP Values) für Merkmalsausprägungen einzelner Zähne bestimmt und veranschaulicht werden. Somit kann der Behandler nachvollziehen, was das Modell zu einer Prognose bewegt hat und weshalb es gerade zu dieser Prognose gekommen ist. Die Merkmalsausprägung beispielsweise für den Knochenabbau kann eine große Heterogenität aufweisen. So tritt nicht bei jedem Zahn Zahnverlust auf, welcher einen relativen röntgenologischen Knochenabbau von $\geq$ 70% (in dieser Studie 35,6%, Tab.2) aufweist. Bei einem klinisch fragwürdig erscheinenden Zahn, könnte ein

Zahnarzt nicht nur eine objektive, erklärbare Entscheidung zur Extraktion (positive Prognose) treffen, sondern würde im Falle des vorhergesagten Zahnerhalts, therapeutische Maßnahmen zur Retention ergreifen können, welche an die individuellen Baustellen des Patienten wie Rauchen oder Knochenabbau angepasst sein könnten. Die Summary Plots ermöglichen ein anschauliches Ranking über die Einflussgrößen der Variablen des gesamten eigenen Patientenstamms und könnten das Augenmerk des Behandlers auf ernst zu nehmende Ausprägungen richten (hoch oder niedrig). Anhand des für die vorliegenden Studie nachuntersuchten Kieler Datensatzes kann man so beispielsweise erkennen, dass die Zahnanzahl zu Beginn der APT, hohe Sondierungstiefen, Kiefer- und Zahntyp sowie ein hoher Knochenverlust großen Einfluss auf die Prognose hat, während das Geschlecht, der Beweglichkeitsgrad des Zahns und teils auch das Rauchverhalten weniger Relevanz zeigen. Die modellspezifischen Rankings zeigen im Rahmen dieser Studie, dass die verschieden komplexen Modelle sich im Umgang mit den Daten unterscheiden. So spielt das Alter bei komplexeren, in unserem Fall weniger verlässlichen, Modellen eine große Rolle (Rang drei und vier ohne SMOTE, Abb. 5c und 6a) und bei der logR eine kleinere (Rang 13, Abb. 5a). Eine bereits erwähnte Studie (Ravida et al., 2019) zeigt, dass das Alter und das Eintreten von Zahnverlust gerade aus parodontalen Gründen, signifikant zusammenhängen. Rang 13 scheint folglich nicht angemessen zu sein. Dies zeigt einerseits, dass für die klinische Anwendbarkeit von SHAP erst einmal ein verlässliches Vorhersagemodell bestehen muss, da erkannte Muster und SHAP Values sonst nicht vertrauenswürdig sind und andererseits, dass SHAP ein nützliches Werkzeug bei der Modellentwicklung darstellt, da es Defizite aufzeigt. Deshalb müssen weitere Bestrebungen unternommen werden, damit diese Interpretationshilfe zukünftig eine brauchbare Ergänzung einer jeden Zahnarztpraxis sein könnte, die die Effizienz und Transparenz von Behandlungen steigern hilft.

## 6.5. Stärken und Limitationen

Wie alle wissenschaftlichen Studien hat auch die vorliegende Studie Stärken und Schwächen, wobei versucht wurde, durch die Anpassung der Methode an die Fragstellung die Einschränkungen so weit möglich zu reduzieren. Zunächst ist die Relevanz des Ziels, Zahnverlust richtig vorhersagen zu können, als Stärke zu nennen. Es spielt aus gesundheitsökonomischen Gründen eine große Rolle und ist auch für die

orale Lebensqualität der Patienten von großer Bedeutung (Bäumer et al., 2018, El Sayed et al., 2019, Tonetti et al., 2017). Alle inkludierten Patienten wurden einheitlich nach demselben Behandlungskonzept behandelt. Extraktionen die während der APT stattfanden, wurden streng überwacht und Zähne, die nach kassenzahnärztlichen Richtlinien oder auch unter einem mehr prothetiklastigem Konzept hätten extrahiert werden müssen, wurden erhalten . So konnten einige Zähne mit fragwürdiger oder hoffnungsloser Prognose während der UPT weiter beobachtet werden, obwohl man Zahnverlust erwartet hätte (Graetz et al., 2011). Außerdem wurde eine erste externe Validierung („Narrow-Validation“) durchgeführt, was eine Einschätzung der Generalisierbarkeit der Modelle ermöglicht. Darüber hinaus wurden zwar in einigen Studien dentale Vorhersagemodelle vorgestellt, teils auch extern validiert (Lindskog et al., 2010, Martinez-Canut et al., 2017, Morelli et al., 2017), aber nach bestem Wissen ist dies die erste Studie in der Zahnmedizin, in der Vorhersagemodelle validiert wurden, die das Problem der Klassenunausgeglichenheit adressieren, bei denen sowohl interne als auch externe Validierung zur Anwendung kamen, bei welcher neben Erkenntnissen aus vorhergehenden Studien der Arbeitsgruppe (Schwendicke et al., 2018) über die besten Entwicklungs- und Validierungsschritte auch bestehende Ansätze zur Berichterstattung, Vertiefung und Verbesserung von Validierungsstudien berücksichtigt und umgesetzt wurden. Allerdings, und das als Limitation dieser Studie, werden in der Literatur prospektive Kohortenstudien zur Entwicklung und Validierung empfohlen sowie bestimmte Verfahren zur Auswahl von Prädiktoren (nicht auf Signifikanzniveau und nicht wie in dieser Studie auf klinischer Verfügbarkeit basierend) gefordert (Cowley et al., 2019). Die Kalibrierungsfähigkeit soll bestimmt werden und das Over-sampling nicht, wie in der Entwicklung der vorliegenden Modelle fälschlicherweise praktiziert, vor der Kreuzvalidierung stattfinden. Außerdem empfehlen die Entwickler von SMOTE die Anwendung dieser Technik in Kombination mit Under-sampling der Majoritätsklasse, was in der zugrundeliegenden Studie nicht erfolgte. Diese Versäumnisse könnten die Leistungen der Vorhersagemodelle negativ beeinflusst haben (Steyerberg et al., 2018). Außerdem treten hohe Bias bei der nachuntersuchten Langzeitkohorte auf, da die Patienten streng selektiert wurden. So müssen die Patienten nicht nur als hoch-compliant eingestuft werden, sondern es wurde auch ein sehr spezifisches konservatives Behandlungskonzept im Rahmen einer universitären Zahnmedizin angewendet (Graetz et al., 2017b) dessen Behandlungsergebnisse nicht verallgemeinert werden dürfen. Es kann angenommen

werden, dass die Anwendung der Modelle in einer zweiten externen Validierung (geforderte „Broad-Validation“) auf eine andere, stark abweichende Kohorte (z.B. mehr Zahnextraktionen oder weniger complainte Patienten (Petit et al., 2019) wohl eine deutlich andere Modellperformance aufweisen würden.

Neben diesen Einschränkungen aufgrund der durchgeführten Maßnahmen, muss weiterhin beachtet werden, dass Raucher beziehungsweise Patienten mit Diabetes mellitus unterrepräsentiert im Verhältnis zu epidemiologischen Untersuchungen vergleichbarer Kohorten mit Parodontitis waren (Volzke et al., 2015). Ebenfalls wurden diese Variablen nicht longitudinal erfasst, was die Konstruktion der statistischen Modelle zwar noch komplexer gestaltet hätte, möglicherweise aber geholfen hätte, einige Ergebnisse hinsichtlich der Gewichtung von Patienten- versus Zahnebenevariablen besser zu verstehen. Obwohl eine große Anzahl an Patienten (n=390) in den Modellen getestet wurde, konnten aufgrund des retrospektiven Studiendesigns nicht alle relevanten patienten- und zahnbezogenen Prädiktoren verwendet und nur wenige Zahnverluste nachuntersucht werden. Ebenso konnten aufgrund fehlender Dokumentation Faktoren, die teils nicht direkt mit der parodontalen Gesundheit verknüpft sind, wie Bruxismus und der restaurative Status, aber ebenfalls für einen Zahnverlust relevant sein könnten (Martinez-Canut, 2015, Martinez-Canut et al., 2017), nicht untersucht werden.

## 7. Ausblick und Praxistauglichkeit

Erstmals wurden in der vorliegenden Dissertationsarbeit Vorhersagemodelle zum Zahnverlust bei Parodontitispatienten sowohl mittels Kreuzvalidierung (intern), als auch extern anhand von Daten aus einer Langzeitkohorte validiert. Das Anwenden dieser Validierungsstrategien war aufgrund der sorgfältig dokumentierten Langzeitdaten einer Kieler-Kohorte von Patienten mit Parodontitis, behandelt in der Klinik für Zahnerhaltung und Parodontologie des Universitätsklinikums Schleswig-Holstein, möglich. Keines der drei validierten Modelle, weder die logistische Regression, der Random Forest noch die Gradient Boosting Machine sind klinisch anwendbar. Die besseren Ergebnisse anhand unserer Testdaten erzielte eine einfache logistische Regression gegenüber der Machine Learning Algorithmen, wobei kein signifikanter Unterschied vorlag. Ein erstmals in diesem Zusammenhang angewendeter Ansatz zum Umgang mit der Klassenunausgeglichenheit der Daten, die Synthetic Minority Over-sampling Technique (SMOTE), eignet sich nicht für diese

Problemstellung. Zukünftige Studien zur Validierung, Entwicklung aber auch zur Weiterentwicklung von Vorhersagemodellen für Zahnverlust sollten das Problem der Klassenunausgeglichenheit fokussieren und dabei weitere Ansätze auf ihre Tauglichkeit hin testen, bestehende Richtlinien befolgen und auf bereits gewonnenen Erkenntnissen aufbauen, anstatt diese zu ignorieren oder zu verwerfen. Hingegen konnte die vorliegende Studie zeigen, dass es sich bei SHapley Additive exPlanations (SHAP) um eine vielversprechende Interpretationshilfe für Prognosen im klinischen Alltag handeln könnte, welche aber die Existenz eines nutzbaren Vorhersagemodells voraussetzt. Diese zu generieren sollte ebenfalls Inhalt zukünftiger Untersuchungen sein.

# 8. Zusammenfassung

Nicht nur in der Medizin, sondern auch in der Zahnmedizin gibt es einen zunehmenden Trend für Vorhersagemodelle zu Folgen irreversibler Entscheidungen (hier z.B. Zahnextraktion). Dafür müssen Modelle anhand eines Datensatzes trainiert (Trainingsdatensatz) und dann anhand von unbekannten Daten getestet (Testdatensatz) werden. Leider weisen die existierenden Modelle zum Zahnverlust methodische Mängel auf, weshalb es das Ziel dieser Arbeit war, drei Vorhersagemodelle (logistische Regression, Random Forest, Gradient Boosting Machine) zum Zahnverlust bei Parodontitis in der Phase der unterstützenden parodontalen Therapie zu validieren und bezüglich ihrer Komplexität zu evaluieren. Ein erstmals auf Daten zum Zahnverlust angewandter Ansatz zum Umgang mit der Klassenunausgeglichenheit (Zahnverlust als seltenes Ereignis), die Synthetic Minority Over-sampling Technique (SMOTE), wurde hinsichtlich einer Verbesserung der Ergebnisse durch Generierung künstlicher Beobachtungen der Klasse Zahnverlust im Trainingsdatensatz getestet und eine Interpretationshilfe von Ergebnissen von Vorhersagemodellen, SHapley Additive exPlanations, wurde für den Gebrauch in der zahnärztlichen Praxis vorgestellt und diskutiert.

Jedes Modell wurde auf die Langzeitdaten einer Kieler-Kohorte von Patienten mit Parodontitis (n=390, 7518 Zähne), welche in der Klinik für Zahnerhaltung und Parodontologie des Universitätsklinikums Schleswig-Holstein über einen Zeitraum von 18,2 ± 5,6 Jahren behandelt wurden, appliziert. Zahnverlust und verschiedene Variablen auf Zahn- und Patientenebene wurden aufgezeichnet. Im Rahmen der „Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis“ Berichterstattungsrichtlinien (2015) und orientiert an bestehenden Erkenntnissen für Validierungsstrategien wurden die Modelle validiert und Metriken der Modellgüte (Area under the Receiver Operating Characteristic Curve (AUC), Accuracy, Sensitivität, Spezifität und der F1-Score) bestimmt, um die Modelle auch unter Verwendung von SMOTE vergleichen zu können.

Zahnverlust war ein seltenes Ereignis (728 von 7518 Zähnen gingen verloren). Ohne SMOTE zeigten alle Modelle eine niedrige Sensitivität, da sie das richtige Prognostizieren der positiven Klasse (tatsächlicher Zahnverlust) nur anhand weniger Beobachtungen erlernen konnten, während eine generell hohe Spezifität die Vorhersagestärke für die viel häufigeren Beobachtungen der negativen Klasse

(Zahnretention) widerspiegelte. Nach der Kreuzvalidierung (intern) ergaben sich für die AUC ohne SMOTE Werte zwischen 0,70 und 0,81 in der externen Validierung. Wobei komplexere Modelle (Random Forest, Gradient Boosting Machine) keine signifikanten Vorteile gegenüber der weniger komplexen logistischen Regression hatten. In der In-sample-Validierung mit SMOTE hingegen kam es zu einer signifikanten Verbesserung der AUC der komplexeren Modelle. Dieses Ergebnis konnte allerdings durch die externe Validierung nicht belegt werden. Unter der Verwendung von SMOTE wurde die Diskriminierungsfähigkeit der komplexeren Modelle in der internen Validierung folglich überschätzt und die Werte der AUC fielen in der externen Validierung von 0,95 und 0,96 auf Werte von 0,64 und 0,7 (Random Forest und Gradient Boosting Machine) ab. Die Anwendung von SHapley Additive exPlanations im klinischen Alltag erfordert die Existenz eines klinisch anwendbaren Vorhersagemodells, könnte aber zukünftig ein Hilfsmittel für Entscheidungen bezüglich personalisierter Behandlungsmaßnahmen und für die Visualisierung für Behandler und Patienten darstellen.

Abschließend muss festgestellt werden, dass keines der drei Vorhersagemodelle wirklich klinisch anwendbar ist und auch SMOTE ungeeignet zur Verbesserung der Diskriminierungsfähigkeit des Zahnverlusts (seltenes Ereignis bei Parodontitis) ist. Deshalb müssen weitere Ansätze getestet und Modelle entsprechend der bestehenden Erkenntnisse erweitert oder rigoros neuentwickelt werden.

# 9. Literaturverzeichnis


Ali, A. S., S.M.; Ralescu, A.L. (2015): Classification with class imbalance problem: a review. *IJASCA,* 7, 176-204.

AlMoharib, H. S., AlMubarak, A., AlRowis, R., Geevarghese, A., Preethanath, R. S.,Anil, S. (2014): Oral Fluid Based Biomarkers in Periodontal Disease: Part 1. Saliva. *Journal of International Oral Health,* 6, 95-103.

AlRowis, R., AlMoharib, H. S., AlMubarak, A., Bhaskardoss, J., Preethanath, R. S.,Anil, S. (2014): Oral fluid-based biomarkers in periodontal disease - part 2. Gingival crevicular fluid. *J Int Oral Health,* 6, 126-35.

American Academy of Periodontology (2008): American Academy of Periodontology statement on risk assessment. *J Periodontol,* 79, 202.

Armitage, G. C. (1999): Development of a classification system for periodontal diseases and conditions. *Ann Periodontol,* 4, 1-6.

Avila, G., Galindo-Moreno, P., Soehren, S., Misch, C. E., Morelli, T.,Wang, H. L. (2009): A novel decision-making process for tooth retention or extraction. *J Periodontol,* 80, 476-91.

Bäumer, A., Kappesz, D., Ozga, A. K., Mertens, C., Eickholz, P.,Pretzl, B. (2018): Oral health-related quality of life and standard of treatment in aggressive periodontitis patients more than 5 years after therapy. *J Clin Periodontol,* 45, 1347-1355.

Blagus, R.,Lusa, L. (2013): SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics,* 14, 106.

Blagus, R.,Lusa, L. (2015): Joint use of over- and under-sampling techniques and cross-validation for the development and assessment of prediction models. *BMC Bioinformatics,* 16, 363.

Bouwmeester, W., Zuithoff, N. P., Mallett, S., Geerlings, M. I., Vergouwe, Y., Steyerberg, E. W., Altman, D. G.,Moons, K. G. (2012): Reporting and methods in clinical prediction research: a systematic review. *PLoS Med,* 9, 1-12.

Branco, P., Torgo, L.,Ribeiro, R. P. (2016): A Survey of Predictive Modelling under Imbalanced Distributions. *ACM Computing Surveys,* 49, 1-31.

Breimann, L. (2001): Random Forests. *Machine Learning,* 45, 5-32.

Chawla, N. V. (2005): Data Mining for Imbalanced Datasets: An Overview. Maimon, O.,Rokach, L., *Data Mining and Knowledge Discovery Handbook*, 1, 853-867, Springer, US.

Chawla, N. V. B., K.W.,Hall, L. O. K., W.P. (2002): SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res,* 16, 321–357.

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y.,Van Calster, B. (2019): A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology,* 110, 12-22.

Collins, G. S., Ogundimu, E. O.,Altman, D. G. (2016): Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat Med,* 35, 214-26.

Cowley, L. E., Farewell, D. M., Maguire, S.,Kemp, A. M. (2019): Methodological standards for the development and evaluation of clinical prediction rules: a review of the literature. *Diagn Progn Res,* 3, 16.

Du, M., Bo, T., Kapellas, K.,Peres, M. A. (2018): Prediction models for the incidence and progression of periodontitis: A systematic review. *J Clin Periodontol,* 45, 1408-1420.
El Sayed, N., Bäumer, A., El Sayed, S., Wieland, L., Weber, D., Eickholz, P.,Pretzl, B. (2019): Twenty years later: Oral health-related quality of life and standard of treatment in patients with chronic periodontitis. *J Periodontol,* 90, 323-330.
Fardal, O., Johannessen, A. C.,Linden, G. J. (2004): Tooth loss during maintenance following periodontal treatment in a periodontal practice in Norway. *J Clin Periodontol,* 31, 550-5.
García, V., Sánchez, J. S.,Mollineda, R. A. (2007): An Empirical Study of the Behavior of Classifiers on Imbalanced and Overlapped Data Sets. *CIARP: Progress in Pattern Recognition, Image Analysis and Applications*. Valparaiso, Chile, November 13-16, 2007, 397-406.
Graetz, C., Dörfer, C. E., Kahl, M., Kocher, T., Fawzy El-Sayed, K., Wiebe, J.-F., Gomer, K.,Rühling, A. (2011): Retention of questionable and hopeless teeth in compliant patients treated for aggressive periodontitis. *J Clin Periodontol,* 38, 707-714.
Graetz, C., Salzer, S., Plaumann, A., Schlattmann, P., Kahl, M., Springer, C., Dorfer, C.,Schwendicke, F. (2017a): Tooth loss in generalized aggressive periodontitis: Prognostic factors after 17 years of supportive periodontal treatment. *J Clin Periodontol,* 44, 612-619.
Graetz, C., Plaumann, A., Schlattmann, P., Kahl, M., Springer, C., Salzer, S., Gomer, K., Dorfer, C.,Schwendicke, F. (2017b): Long-term tooth retention in chronic periodontitis - results after 18 years of a conservative periodontal treatment regimen in a university setting. *J Clin Periodontol,* 44, 169-177.
Gursoy, U. K., Könönen, E., Pussinen, P. J., Tervahartiala, T., Hyvarinen, K., Suominen, A. L., Uitto, V. J., Paju, S.,Sorsa, T. (2011): Use of host- and bacteria-derived salivary markers in detection of periodontitis: a cumulative approach. *Dis Markers,* 30, 299-305.
Hamp, S. E., Nyman, S.,Lindhe, J. (1975): Periodontal treatment of multirooted teeth. Results after 5 years. *J Clin Periodontol,* 2, 126-35.
Hastie, T. T., R. Friedman, J. (2009): The Elements of Statistical Learning: Data mining, Inference, and Prediction, 2, New York: Springer, Stanford, California.
Japkowicz, N., Chawla N., Kolcz, A. (2003): Editorial: Special Issue on Learning from Imbalanced Data Sets. *Sigkdd Explorations,* 6, 1.
Kassebaum, N. J., Bernabe, E., Dahiya, M., Bhandari, B., Murray, C. J.,Marcenes, W. (2014): Global Burden of Severe Periodontitis in 1990-2010: A Systematic Review and Meta-regression. *J Dent Res,* 93, 1045-1053.
Kassebaum, N. J., Smith, A. G. C., Bernabé, E., Fleming, T. D., Reynolds, A. E., Vos, T., Murray, C. J. L., Marcenes, W.,Collaborators, G. B. D. O. H. (2017): Global, Regional, and National Prevalence, Incidence, and Disability-Adjusted Life Years for Oral Conditions for 195 Countries, 1990-2015: A Systematic Analysis for the Global Burden of Diseases, Injuries, and Risk Factors. *J Dent Res,* 96, 380-387.
Kattan, M. W., Yu, C., Stephenson, A. J., Sartor, O.,Tombal, B. (2013): Clinicians versus nomogram: predicting future technetium-99m bone scan positivity in patients with rising prostate-specific antigen after radical prostatectomy for prostate cancer. *Urology,* 81, 956-61.
Kim, C. K., Choi, S. H., Kim, T. S., Kaltschmitt, J.,Eickholz, P. (2006): The infrabony defect and its determinants. *J Periodontal Res,* 41, 498-502.

Krois, J. G., C., Holtfreter, B., Brinkmann, P., Kocher, T.,Schwendicke, F. (2019): Evaluating Modeling and Validation Strategies for Tooth Loss. *J Dent Res,* 98, 1088–1095.
Kundu, S., Mazumdar, M.,Ferket, B. (2017): Impact of correlation of predictors on discrimination of risk models in development and external populations. *BMC Med Res Methodol,* 17, 63.
Lang, N. P.,Tonetti, M. S. (2003): Periodontal risk assessment (PRA) for patients in supportive periodontal therapy (SPT). *Oral Health Prev Dent,* 1, 7-16.
Lindhe, J.,Nyman, S. (1977): The role of occlusion in periodontal disease and the biological rationale for splinting in treatment of periodontitis. *Oral Sci Rev,* 10, 11-43.
Lindskog, S., Blomlof, J., Persson, I., Niklason, A., Hedin, A., Ericsson, L., Ericsson, M., Jarncrantz, B., Palo, U., Tellefsen, G., Zetterstrom, O.,Blomlof, L. (2010): Validation of an algorithm for chronic periodontitis risk assessment and prognostication: risk predictors, explanatory values, measures of quality, and clinical use. *J Periodontol,* 81, 584-93.
Lundberg, S.,Lee, S. (2017): A unified Approach to interpreting Model Predictions. *NIPS*. Long Beach Convention Center, Long Beach,
Martinez-Canut, P. (2015): Predictors of tooth loss due to periodontal disease in patients following long-term periodontal maintenance. *J Clin Periodontol,* 42, 1115-25.
Martinez-Canut, P., Alcaraz, J., Alcaraz, J., Jr., Alvarez-Novoa, P., Alvarez-Novoa, C., Marcos, A., Noguerol, B., Noguerol, F.,Zabalegui, I. (2017): Introduction of a prediction model to assigning periodontal prognosis based on survival time. *J Clin Periodontol,* 45, 46-55.
McGuire, N. (1996): Prognosis versus actual outcome. III. The effectiveness of clinical parameters in accurately predicting tooth survival. . *J Periodontol,* 67, 666-674.
Menon, A. K., Agarwal, H. N. S.,Chawla, S. (Year): On the Statistical Consistency of Algorithms for Binary Classification under Class Imbalance. Proceedings of the 30th International Conference on Machine Learning, Cambridge, MA, *PMLR*, 28, 603-611.
Moons, K. G., Altman, D. G., Reitsma, J. B., Ioannidis, J. P., Macaskill, P., Steyerberg, E. W., Vickers, A. J., Ransohoff, D. F.,Collins, G. S. (2015): Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med,* 162, W1-73.
Morelli, T., Moss, K. L., Beck, J., Preisser, J. S., Wu, D., Divaris, K.,Offenbacher, S. (2017): Derivation and Validation of the Periodontal and Tooth Profile Classification System for Patient Stratification. *J Periodontol,* 88, 153-165.
Naseriparsa, M.,Kashani, M. M. R. (2013): Combination of PCA with SMOTE Resampling to Boost the Prediction Rate in Lung Cancer Dataset. *Int J Comput Appl,* 77, 33-38.
Nunn, M. E., Fan, J., Su, X., Levine, R. A., Lee, H.-J.,McGuire, M. K. (2012): Development of prognostic indicators using classification and regression trees for survival. *Periodontol 2000,* 58, 134-142.
Ochani, M., Sawarkar, S. D.,Narwane, S. (2019): A novel approach to handle class imbalance : A Survey. *International Journal of Engineering Development and Research,* 7, 419-422.

Olivera, A. R., Roesler, V., Iochpe, C., Schmidt, M. I., Vigo, A., Barreto, S. M.,Duncan, B. B. (2017): Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes - ELSA-Brasil: accuracy study. *Sao Paulo Med J,* 135, 234-246.

Perlich, C., Provost, F.,Simonoff, J. S. (2003): Tree Induction vs. Logistic Regression: A Learning-Curve Analysis. *J Mach Learn Res,* 4, 211-255.

Petit, C., Schmeltz, S., Burgy, A., Tenenbaum, H., Huck, O.,Davideau, J. L. (2019): Risk factors associated with long-term outcomes after active and supporting periodontal treatments: impact of various compliance definitions on tooth loss. *Clin Oral Investig,* 23, 4123-4131.

Ramseier, C. A., Anerud, A., Dulac, M., Lulic, M., Cullinan, M. P., Seymour, G. J., Faddy, M. J., Burgin, W., Schatzle, M.,Lang, N. P. (2017): Natural history of periodontitis: Disease progression and tooth loss over 40 years. *J Clin Periodontol,* 44, 1182-1191.

Ravida, A., Qazi, M., Troiano, G., Saleh, M. H. A., Greenwell, H., Kornman, K.,Wang, H. L. (2019): Using periodontal staging and grading system as a prognostic factor for future tooth loss: A long-term retrospective study. *J Periodontol,* 00, 1-8.

Schmalz, G., Li, S., Burkhardt, R., Rinke, S., Krause, F., Haak, R.,Ziebolz, D. (2016): MicroRNAs as Salivary Markers for Periodontal Diseases: A New Diagnostic Approach? *Biomed Res Int,* 2016, 1-14.

Schwendicke, F., Stolpe, M.,Graetz, C. (2017): Cost comparison of prediction-based decision-making for periodontally affected molars. *J Clin Periodontol,* 44, 1145-1152.

Schwendicke, F., Graetz, C., Stolpe, M.,Dorfer, C. E. (2014): Retaining or replacing molars with furcation involvement: a cost-effectiveness comparison of different strategies. *J Clin Periodontol,* 41, 1090-7.

Schwendicke, F., Schmietendorf, E., Plaumann, A., Salzer, S., Dorfer, C. E.,Graetz, C. (2018): Validation of multivariable models for predicting tooth loss in periodontitis patients. *J Clin Periodontol,* 45, 701-710.

Steyerberg, E. W.,Vergouwe, Y. (2014): Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European heart journal,* 35, 1925-1931.

Steyerberg, E. W.,Harrell, F. E., Jr. (2016): Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol,* 69, 245-247.

Steyerberg, E. W., Uno, H., Ioannidis, J. P. A.,van Calster, B. (2018): Poor performance of clinical prediction models: the harm of commonly applied methods. *J Clin Epidemiol,* 98, 133-143.

Sun, Y., Wong, A. K. C.,Kamel, M. S. (2011): Classification of Imbalanced Data: A Review. *International Journal of Pattern Recognition and Artificial Intelligence,* 23, 687-719.

Toll, D. B., Janssen, K. J., Vergouwe, Y.,Moons, K. G. (2008): Validation, updating and impact of clinical prediction rules: a review. *J Clin Epidemiol,* 61, 1085-94.

Tonetti, M. S., Greenwell, H.,Kornman, K. S. (2018): Staging and grading of periodontitis: Framework and proposal of a new classification and case definition. *J Periodontol,* 89, 159-172.

Tonetti, M. S., Jepsen, S., Jin, L.,Otomo-Corgel, J. (2017): Impact of the global burden of periodontal diseases on health, nutrition and wellbeing of mankind: A call for global action. *J Clin Periodontol,* 44, 456-462.

Vergouwe, Y., Steyerberg, E., Eijkemans, M.,Habbema, J. (2002): Validity of prognostic models: When is a model clinically useful? *Semin Urol Oncol,* 20, 96-107.
Vickers, A. J.,Elkin, E. B. (2006): Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making,* 26, 565-74.
Volzke, H., Ittermann, T., Schmidt, C. O., Baumeister, S. E., Schipf, S., Alte, D., Biffar, R., John, U.,Hoffmann, W. (2015): Prevalence trends in lifestyle-related risk factors. *Dtsch Arztebl Int,* 112, 185-92.
Wallert, J., Tomasoni, M., Madison, G.,Held, C. (2017): Predicting two-year survival versus non-survival after first myocardial infarction using machine learning and Swedish national register data. *BMC Med Inform Decis Mak,* 17, 99.
Wedderburn, J. A. N. R. W. M. (1972): Generalized Linear Models. *Journal of the Royal Statistical Society,* 135,No. 3, 370-384.
Weiss, G. P., F (2003): Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. *Journal of Artificial Intelligence Research,* 19, 315-354.
Winkelhoff, A. J. V., Rams, T. E.,Slots, J. (1996): Systematic antibiotic therapy in periodontics. *Periodontol 2000,* 10, 45-78.
Yen, S.-J.,Lee, Y.-S. (2006): Under-Sampling Approaches for Improving Prediction of the Minority Class in an Imbalanced Dataset. Huang, D.-S., Li, K.,Irwin, G. W., *Intelligent Control and Automation*, 731-740, Springer-Verlag, Berlin Heidelberg.

# 10. Anhang

**MEDIZINISCHE FAKULTÄT**
**DER CHRISTIAN-ALBRECHTS-UNIVERSITÄT ZU KIEL**

ETHIK-KOMMISSION

A.D. 1665

Universitäts-Kinderklinik · Schwanenweg 20 · 24105 Kiel

Postadresse:
Arnold-Heller-Straße 3 / Haus 9
D-24105 Kiel

Telefon 04 31 / 597-18 09
Telefax 04 31 / 597-53 33
ethikkomm@email.uni-kiel.de

Dr. Christian Graetz
Klinik für Zahnerhaltung und Parodontologie
UKSH, Campus Kiel
24105 Kiel

Datum: 14. August 2013

| | |
|---|---|
| **AZ:** | **D 489/13** (bitte stets angeben) |
| **Studienplan:** | **Retrospektive Studie zum Langzeiterhalt von Zähnen und Implantaten bei aggressiver, chronischer Parodontitis und parodontalgesunden Patienten der Klinik für Zahlerhaltungskunde und Parodontologie sowie der Klinik für Mund-Kiefer und Gesichtschirurgie des Universitätsklinikums Schleswig-Holstein, Campus Kiel** |
| **In Bezug auf Studienplan:** | **D 442/10** **Retrospektive Studie zum Langzeiterhalt von Zähnen bei aggressiver und chronischer Parodontitis** |
| **Antragsteller:** | **Dr. Christian Graetz** |
| **Studienleiter:** | **Prof. Dr. C. Dörfer** |
| **Antrag vom:** | **2. August 2013 (Eingang 7. August 2013)** Anschreiben, Studienplan |

Sehr geehrter Herr Kollege Graetz,

vielen Dank für den obengenannten Antrag zur Beratung gemäß § 15 Berufsordnung (BO) der Ärztekammer Schleswig-Holstein.

Nach Durchsicht der Unterlagen durch die Geschäftsstelle und durch mich als Vorsitzenden der Ethik-Kommission bestehen gegen die Durchführung der Studie keine berufsethischen und berufsrechtlichen Bedenken.

Mit freundlichen kollegialen Grüßen

Prof. Dr. med. H. M. Mehdorn
Vorsitzender der Ethik-Kommission

Dr. med. Christine Glinicke
Geschäftsführung der Ethik-Kommission

**MEDIZINISCHE FAKULTÄT**
**DER CHRISTIAN-ALBRECHTS-UNIVERSITÄT ZU KIEL**

ETHIK-KOMMISSION

A.D. 1665

Universitäts-Kinderklinik · Schwanenweg 20 · 24105 Kiel

PD Dr. Christian Graetz
UKSH Campus Kiel
Klinik für Zahnerhaltungskunde und
Parodontologie
Arnold-Heller-Straße 3, Haus 26
24105 Kiel

Postadresse:
Arnold-Heller-Straße 3 / Haus 9
D-24105 Kiel

Telefon 0431 / 500-14191
Telefax 0431 / 500-14195
ethikkomm@email.uni-kiel.de

Datum: 24. November 2017

| | |
|---|---|
| **AZ.:** | **D 566/17** (bitte stets angeben) |
| **Studienplan:** | **Untersuchung des Bindungsverhaltens von Parodontitispatienten in Bezug auf die Auflösung des Behandlungsverhältnisses in der initialen, aktiven sowie unterstützenden Parodontitistherapie bei Patienten mit aggressiven und chronischen Parodontalerkrankungen in der Klinik für Zahnerhaltung und Parodontologie des UKSH, Campus Kiel**<br>Anschreiben v. 13.11.17; Prüfplan Version 2017-11; Patienteninformation Version 2017-11; Fragebogen |
| **In Bezug auf:** | **D 442/10, D 489/13** |
| **Studienleiter und Antragsteller:** | **PD Dr. Christian Graetz, Rebecca Kowalski, UKSH Campus Kiel** |
| **Antrag vom:** | **13. November 2017 (Eingang: 16. November 2017)** |

Sehr geehrter Herr Dr. Graetz,

wir bestätigen den Eingang des obengenannten Antrages zur Beratung gemäß § 15 Berufsordnung (BO) der Ärztekammer Schleswig-Holstein. Nach Durchsicht der Unterlagen durch die Geschäftsstelle und durch mich als Vorsitzenden der Ethik-Kommission bestehen gegen die Durchführung der Studie keine berufsethischen und berufsrechtlichen Bedenken.
**Die im Folgenden aufgeführten Hinweise müssen beachtet werden:**
1. Bitte überarbeiten Sie die Einwilligungserklärung (es liegt nur ein Datenschutzpassus vor- gibt es Auftraggeber und Kooperationspartner? Ansonsten bitte löschen. Es s fehlt die Arztunterschrift etc.) gemäß den Mustertexten unter *http://www.medizin.uni-kiel.de/de/fakultaet/ethik-kommission/sonstige-studien* und **reichen Sie eine entsprechend geänderte Version per E-Mail nach.**
2. Es wird darauf hingewiesen, dass künftige Änderungen der Studie der Ethik-Kommission anzuzeigen sind und gegebenenfalls eine erneute Beratung erforderlich machen.
3. Die ethische und rechtliche Verantwortung für die Durchführung dieser Studie verbleibt beim Studienleiter.
4. Gemäß Deklaration von Helsinki **muss** der Ethik-Kommission nach Studienende ein Abschlussbericht vorgelegt werden, der eine Zusammenfassung der Ergebnisse und Schlussfolgerungen der Studie enthält.

Wir wünschen Ihnen für die Durchführung Ihrer Studie viel Erfolg.

Mit freundlichen kollegialen Grüßen

Prof. Dr. med. H. M. Mehdorn
Vorsitzender der Ethik-Kommission

Dr. med. Christine Glinicke
Geschäftsführung der Ethik-Kommission

# 11. Danksagung

An dieser Stelle möchte ich allen Personen danken, die mich bei der Anfertigung meiner Dissertation unterstützt haben.

Mein besonderer Dank gilt Herrn Dr. Graetz für die ausgezeichnete Betreuung während der Umsetzung der Arbeit und vor allem zum Ende hin, wo ein geschulter Blick fürs Detail und konsequent schnelle Hilfestellungen eine enorme Unterstützung für mich darstellten.

Außerdem möchte ich mich bei Herrn Prof. Schwendicke für die Möglichkeit der Teilnahme an dem Projekt bedanken und natürlich bei Herrn Dr. Krois, der mir jederzeit zuverlässig mit Rat und vor allem mit Tat beigestanden hat.

Abschließend danke ich bei dieser Gelegenheit noch meiner Mutter für ihre unglaubliche Geduld mit mir während des Studiums und in der aktiven Schreibphase der Dissertation. Auf dich ist immer Verlass! Und: Es ist vorüber! Danke!